ANNOTATION AND CHARACTERIZATION OF CLASS 2 TRANSPOSABLE ELEMENTS IN CEREAL GRASS GENOMES

by

YUJUN HAN

(Under the Direction of Susan R Wessler)

ABSTRACT

With the improvement in DNA sequencing technology, more and more species have and will be sequenced. To date, one of the amazing genomic discoveries is that the genomes of most higher eukaryotes are largely composed of sequences derived from transposable elements (TEs). My research interests are related to Class 2 (DNA) TEs and especially one type called MITEs. MITEs have been hypothesized to be important players in gene regulation and genome evolution because of their high copy number and preferential association with genes. The availability of genomic sequences from many plant species provides the raw material to determine the location of TEs relative to plant genes at the genome level. To accomplish this requires computer programs that can accurately discovery TEs from large genomic databases. Although numerous TE annotation programs have been developed, none are very successful at both identifying and characterizing new Class 2 elements. For this reason I developed my own tools. This dissertation is composed of four chapters. Chapter 1 provides an introduction to the TE classification system and currently available TE discovery algorithms and programs. Chapter 2 describes a multifunctional pipeline named TARGeT (Tree Analysis of Related Genes and Transposons) that can use either a DNA or a protein sequence as the query to identify, retrieve and characterize homologs from DNA sequence databases. Chapter 3 describes MITE-Hunter, a TE discovery program that can find small nonautonomous DNA TEs (especially MITEs) in genomic datasets. MITE-Hunter was evaluated by applying it to the rice genome and comparing its results

with rice TE databases as well as results generated by similar programs. Chapter 4 describes DNA TEs annotated using TARGeT and MITE-Hunter and their distribution in sequenced genomes from the grass clade. Of greatest interest is that MITEs cluster in the promoters of a large fraction of annotated genes, and are especially enriched within 1 Kb of the transcription start site. Putative mechanisms for the observed enrichment and the potential impact of MITEs on the evolution of gene expression are discussed.

INDEX WORDS: Annotation characterization class 2 nonautonomous TEs MITEs plant genes

ANNOTATION AND CHARACTERIZATION OF CLASS 2 TRANSPOSABLE ELEMENTS IN CEREAL GRASS GENOMES

by

YUJUN HAN

B.S., Nankai University, 2001

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2010

ANNOTATION AND CHARACTERIZATION OF CLASS 2 TRANSPOSABLE ELEMENTS IN CEREAL GRASS GENOMES

by

YUJUN HAN

Major Professor:	Susan R Wessler
Committee:	R. Kelly Dawe
	Katrien Devos
	Russell L. Malmberg
	Jim Leebens-Mack

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2010

ACKNOWLEDGEMENTS

It has been my great fortune to be a student of the Plant Biology department at the University of Georgia, where I had a very wonderful graduate experience. I greatly appreciate the support, guidance and inspiration from my major advisor Susan Wessler and my committee members, Kelly Dawe, Katrien Devos, Russell Malmberg and Jim Leebens-Mack. I am grateful to my colleagues and friends Nathan Hancock, James Burnette, Yaowu Yuan, Aaron Richardson, Dawn Holligan, Ken Naito, Eleanor Kuntz, Han Zhang, Lin Guo, Xiaoyu Zhang, Ning Jiang, Feng Zhang, Eunyoung Cho, Guojun Yang and Christiane Bradshaw for their help and friendship. I sincerely thank department staff Susan Watkins and Carla Ingram for their administrative assistance. Finally, I deeply indebted to my mother Aifen Yu, my father Chengyang Han and my wife Shanshan Qin for their never changing support, encouragement and love.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

TEs have been classified into two classes based on whether RNA or DNA is the transposition intermediate: class 1 (RNA) TEs and class 2 (DNA) TEs. class 2 TEs have been further divided into superfamilies based on the phylogeny of the encoded transposase gene (Wicker *et al.* 2007). To date, six class 2 superfamilies have been found to be widely distributed in plants: *Tc1*/*mariner*, *PIF*/*Harbinger*, *hAT*, *Mutator*, *CACTA* and *Helitron* (Feschotte and Pritham 2007; Wicker *et al.* 2007). With the exception of *Helitrons*, all superfamilies have terminal inverted repeats (TIRs) and transpose through a “cut and paste” mechanism.

TEs can also be classified into autonomous and nonautonomous elements based on whether or not they encode functional transposase. Some nonautonomous elements contain transposase genes that have accumulated mutations and are now inactive. In contrast, miniature inverted repeat transposable elements (MITEs) are a special class of nonautonomous element that do not encode transposase (Bureau and Wessler 1992; Bureau and Wessler 1994). In addition, unlike other nonautonomous TE families that typically have few copies (< 50) of different length, MITE families are characterized by their short length (usually less than 500bp), high copy number and structural homogeneity.

Class 2 superfamilies in plants

The *hAT* superfamily. The first discovered TEs were *Activator* and *Dissociation* (*Ac/Ds*), which are now classified as members of the *hAT* superfamily. *hAT* was named after the first three elements found in this superfamily: *hobo* from the fruit fly *Drosophila melangaster* (Calvi *et al.* 1991), *Ac/Ds* from *Zea Mays* (McClintock 1947; McClintock 1948) and *Tam3* from *Antirrhinum majus* (Feldmar and Kunze 1991; Hehl *et al.* 1991). The *hAT* superfamily is widely distributed in plants (Frank *et al.* 1997; Fujino *et al.* 2009), fungi (Kempken and Kuck 1996), worms (Bigot *et al.* 1996), fish (Koga *et al.* 1996) and human(Esposito *et al.* 1999). Many active *hATs* have been reported, however, no active *hAT* element has been found in mammals. A recently discovered nonautonomous *hAT* in the bat may be the first but it awaits further characterization (Ray *et al.* 2007).

The structure of *hAT* elements is simple: they have TIRs (most 5 to 27 bps) flanking a single transposase gene. The transposase makes an 8 bp staggered cut of (largely random) host sequence. Upon insertion, an 8 bp target site duplication (TSD) of host sequence is generated, which can be exploited in *hAT* element identification. To date, only a few MITE-like families have been reported in this superfamily (Holyoake and Kidwell 2003; Moreno-Vazquez *et al.* 2005).

The *Tc1/mariner* Superfamily. The *Tc1*/*mariner* superfamily has been found in numerous species of fungi, plants and animals, and is perhaps the most widely dispersed TE superfamily (de Queiroz and Daboussi 2003; Hartl *et al.* 1997). In 1983, the 1.7 kb transposon *Tc1* was discovered in *C. elegans* (Emmons *et al.* 1983). In 1985,

the analysis of an unstable mutation in *D. mauritiana* led to the discovery of the 1.3 kb *mariner* element (Jacobson and Hartl 1985; Jacobson *et al.* 1986). Based on structural and sequence comparisons, *Tc1* and *mariner* were grouped into a superfamily named *Tc1/mariner* (Hartl *et al.* 1997; Plasterk 1996). In 1992, the third member of this superfamily was identified in *D. mauritiana* and was named *pogo* (Tudor *et al.* 1992). *Tc1*, *mariner* and *pogo* are believed to be monophyletic in origin (Capy *et al.* 1996). This superfamily also has a close relative in bacteria called IS630 (Henikoff 1992), and, as such, this superfamily is also called *IS630-Tc1-mariner*.

The *Tc1*/*mariner* superfamily has been well studied and its copy number has been found to vary tremendously among species. The copy number of *Tc1*/*mariner* elements in sequenced plant genomes is relatively low (about 50 copies). In rice, *mariner*-like transposons have been demonstrated to encode the transposase necessary for amplifying *Stowaway* MITEs (Feschotte *et al.* 2005; Feschotte *et al.* 2003; Yang *et al.* 2009).

The structure of *Tc1*/*mariner* elements is also simple. TEs in this superfamily are short (< 5 Kb) and have short TIRs (~20 to 30 bp). Autonomous TEs encode a single gene, the transposase, which has a N-terminal domain with a HTH motif that binds to the TIRs of the encoding element (Lampe *et al.* 2001; Pietrokovski and Henikoff 1997; Zhang *et al.* 2001a) and has a C- terminal domain with the “DDE/D” motif that catalyzes DNA cleavage and strand transfer during the transposition (Doak *et al.* 1994; Hartl *et al.* 1997). Upon insertion, *Tc1*/*mariner* elements generate a TSD that is always the dinucleotide TA.

The *PIF/Harbinger* Superfamily . The *PIF/Harbinger* superfamily was identified more recently compared to other Class 2 TE superfamilies. Because of its recent discovery, there is not yet an accepted name for the superfamily. In this dissertation it is called *PIF/Harbinger*, as this reflects the names of two founding member. *Harbinger* was first recognized by computational analysis from *Arabidopsis* (Kapitonov and Jurka 1999) while the first actively transposing nonautonomous members were isolated from maize and named *P Instability Factor* (*PIF*) (Zhang *et al.* 2001b). Finally, the first active autonomous members, *Ping* and *Pong* were isolated from rice (Jiang *et al.* 2003). The *PIF* and *Ping/Pong* elements define the two clades that make up this superfamily. The *PIF* and *Ping/Pong* clades are distantly related to the bacterial IS5 group. TEs of this superfamily have been found in plants, insects, fungi, fish and bacteria (Grzebelus and Simon 2009; Kapitonov and Jurka 2004; Zhang *et al.* 2001b) but not in mammals. Autonomous *PIF/Harbinger* TEs are characterized by short TIRs (~30 bp) and are flanked by 3 bps TSDs (TAA/TTA, mostly). Unlike most other Class 2 TEs, members of this superfamily encode two open reading frames (ORFs) (Kapitonov and Jurka 2004; Zhang *et al.* 2004). One ORF encodes the transposase that has a DDE/D domain and the other encodes a protein containing a *myb*-like DNA binding domain that is necessary for transposition (Hancock *et al.* ; Jiang *et al.* 2003).

The *PIF/Harbinger* superfamily is perhaps best known for its association with *Tourist* MITEs. The first characterized active MITE, a 430-bp Tourist element named mPing, was identified through sequence analysis in rice and showed high activity in rice cell culture and in some recently domesticated rice strains (Jiang *et al.* 2003; Naito *et al.*

2006). Ping and Pong have been shown to be the autonomous elements that catalyze the transposition of mPing (Yang *et al.* 2007). In maize, an active PIF element named PIFa has been shown to be the autonomous element of a Tourist-like MITE family called miniature PIF (mPIF) (Zhang *et al.* 2001b).

The *Mutator* superfamily. The first *Mutator* TE (*Mu* element) was found by Donald Robertson in a line of maize with an unusually high mutation rate (Robertson 1978). *MuDR* elements were later identified as the autonomous elements of the nonautonomous *Mu* elements in maize (Chomet *et al.* 1991; Hershberger *et al.* 1991; Qin *et al.* 1991). *MuDR* has two genes, *mudrA* and *mudrB*. *mudrA* encodes the transposase and *mudrB* encodes a protein that is required for integration (Hershberger *et al.* 1995; Lisch *et al.* 1999). Importantly, *mudrA* homologs have been identified in plants, animals, fungi and bacteria (Chalvet *et al.* 2003; Eisen *et al.* 1994; Lisch 2002; Makarova *et al.* 2002), but *mudrB* only exists in maize and its close relatives (Lisch 2002; Lisch *et al.* 2001). Besides *MuDR*, three other families of *Mu*-like elements (MULEs) have been identified in maize: *Jittery*, TRAP and TAFT (Comelli *et al.* 1999; Wang and Dooner 2006; Xu *et al.* 2004). Recently, new families of MULEs called non-TIR MULEs that have very short TIRs were identified in both plants and in yeast (Neuveglise *et al.* 2005; Yu *et al.* 2000).

Mu elements have long TIRs (~200 bp) that are usually flanked by 8-10 (usually 9) bp TSDs. One unusual feature of the *Mutator* superfamily is that different nonautonomous *Mu* families have similar TIRs, but the internal sequences are unique to each family and are believed to originate from captured or transduplicated sequences

(Chandler *et al.* 1986; Lisch 2002). Such a transduplication mechanism is believed the cause of a special kind of *Mu* element called *Pack-MULE*s that capture genes from their host (Holligan *et al.* 2006; Jiang *et al.* 2004). To date, very few MITE-like TEs have been reported in this superfamily (Kuang *et al.* 2009).

The *CACTA* superfamily. The first and also the best characterized *CACTA* transposon is *En*/*Spm* that was independently discovered in 1953 (Peterson 1953) and 1954 (McClintock 1954). The *CACTA* superfamily has been reported mainly in plants (Langdon *et al.* 2003; Tian 2006). Recently relatives of *CACTA* were reported in animals and fungi (DeMarco *et al.* 2006). *CACTA* elements are characterized by the sequence "CACTA/G" at the ends and short TIRs (~13 bp). In addition, they generate a 3 bp TSD. A *CACTA* autonomous element encodes two proteins that are expressed from a single alternatively spliced transcription unit (Chopra *et al.* 1999; Masson *et al.* 1991). Both proteins as well as the TIRs and a minimal number of target motifs within the element are needed to form the active transposition complexes (Frey *et al.* 1990). Gene capture events have been reported by *CACTA* elements in plants (Alix *et al.* 2008; Zabala and Vodkin 2007; Zabala and Vodkin 2005). No MITE-like TEs have been reported in this superfamily.

TE annotation programs

Although many TE annotation programs have been developed, their algorithms can be classified into four main groups: 1) homology based, 2) structure based, 3) polymorphism based, and 4) *de novo* repeat discovery. Programs using these different approaches are discussed in the section below.

Homology based programs. These programs detect TEs by their sequence similarity to known elements. Theoretically, any sequence pairwise alignment tool can be used for homology based TE annotation, such as BLAST (Altschul *et al.* 1997), FASTA (Pearson 1990) and CrossMatch (unpublished, http://www.genome.washington.edu). To make it easier for the user, several program packages have been developed specifically for the purpose of TE annotation. The best known is RepeatMasker (unpublished, http://repeatmasker.org) which uses Rebase (Jurka 2000) as its library and CrossMatch or WU-BLAST (http://blast.wustl.edu/) as its alignment tool. Other program packages such as MaskerAid (Bedell *et al.* 2000) and GPS (McClure *et al.* 2005) are functionally similar with the main difference being the utilization of different pairwise alignment tools.

Compared to other TE annotation methods, homology based TE annotation programs have a higher specificity. When analyzing a newly sequenced genome with no available TE information, a homology based TE annotation method must use TE sequences from other (ideally closely) related species. Because the TE sequences often diverge rapidly even between closely related species, using DNA sequences of known TEs as the query can result in low sensitivity. Alternatively, use of a translated

alignment algorithm like TBLASTN with the conserved regions from transposase sequences as the query can dramatically increase the chances of detecting new TEs. However, this approach can only find the coding regions of autonomous TEs. Nonautonomous elements that miss coding regions cannot be found in this way. In conclusion, homology based TE annotation approaches are powerful at detecting previously characterized TEs but are much less effective at discovering divergent or uncharacterized elements from newly sequenced genomes.

***De novo* repeat discovery approach.** In the last twenty years large genomes have been sequenced at an increasing rate with the aid of improved DNA sequencing techniques, increased computing power and better algorithms. To annotate a newly sequenced genome it is imperative to first annotate and mask the repeats. As mentioned above, homology based TE annotation methods are not efficient for such an application. Therefore *de novo* repeat annotation programs have been developed and widely used. There are many different *de novo* repeat discovery programs, such as Reputer (Kurtz and Schleiermacher 1999), RepeatMatch (Delcher *et al.* 1999), RepeatFinder (Volfovsky *et al.* 2001), Recon (Bao and Eddy 2002), RepeatGluer (Pevzner *et al.* 2004) and RepeatScout (Price *et al.* 2005). Although using different algorithms, these programs search for repeats based on one basic feature: they are repetitive. There are two ways to find repetitive sequences: all by all sequence alignment (such as Recon) or searching for short sequences with high sequence identity followed by extension and connection (such as RepeatScout). When searching for new repeats in a large genome, the latter approach is much faster.

Although programs using *de novo* approaches are powerful at discovering new TEs, they have the following limits. First, they are not good at finding low copy TEs, especially in draft sequence. In comparison, homology-based TE finding programs have no such limitation because even a single copy element can be found as long as it has enough similarity to the query sequence. Second, it is difficult for programs using the *de novo* approach to find full length TEs. This is because draft genome sequences have many TE fragments and TEs can insert into each other, making it difficult to find full-length TEs. Finally and most importantly, different kinds of repetitive sequences such as tandem repeats, gene families and TEs mix together in the output of *de novo* programs. To obtain a high resolution TE annotation result, a great deal of time consuming manual effort is required to check, characterize and classify potential TEs. This problem can be solved, but only in part by the use of the program PILER (Edgar and Myers 2005) which can only separate dispersed repeats from tandem repeats.

2.3 Structure-based TE annotation. Some TE families can be identified through structure-based TE annotation approaches because they have unique structural features that can be recognized by computer programs. To date there are several kinds of TE families that can be discovered in this way such as MITEs and LTR retrotransposons. Most structure based programs were developed to find LTR elements such as LTR_FINDER (Xu and Wang 2007), LTR_STRUC (McCarthy and McDonald 2003) and LTRharvest (Ellinghaus *et al.* 2008).

One big advantage of structure-based TE discovery approaches is that they do not rely on previously known sequences or the repetitive feature of TEs. As such, they

can be applied to newly sequenced genomes and they can find low copy elements. A limitation of such TE discovery approaches is that they cannot find TE families that lack significant structural features. Second, not all sequences that satisfy the structural feature of a certain TE family are real TEs, which can result in high false positive rates. Two structure-based TE discovery programs, FINDMITE (Tu 2001) and MUST, are currently available for finding new MITEs but both have high false positive rates where significant manual curation is required.

2.4 Polymorphism-based TE annotation. The activity of TEs can result in genomic sequence polymorphisms (insertion or excision) that can be detected as "empty sites" when sequences are compared. This approach has been used to identify TEs from several *Drosophila* genomes (Caspi and Pachter 2006). The advantage of such an approach is that it doesn't rely on sequence or structural similarity, and it is easy to determine the boundaries of TEs. However, this approach has several limitations. First, it can only be performed between very closely related species, which is perhaps the biggest drawback. Second, it faces the same difficult question as the *de novo* TE annotation approach: how to classify and characterize the elements that it finds. Finally, no program of this kind is publicly available.

References

Alix, K., Joets, J., Ryder, C.D., Moore, J., Barker, G.C., Bailey, J.P., King, G.J., and Pat Heslop-Harrison, J.S. 2008. The CACTA transposon Bot1 played a major role in Brassica genome divergence and gene proliferation. *Plant J* **56:** 1030-1044.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Bao, Z. and Eddy, S.R. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* **12:** 1269-1276.

Bedell, J.A., Korf, I., and Gish, W. 2000. MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics* **16:** 1040-1041.

Bigot, Y., Auge-Gouillou, C., and Periquet, G. 1996. Computer analyses reveal a hobo-like element in the nematode Caenorhabditis elegans, which presents a conserved transposase domain common with the Tc1-Mariner transposon family. *Gene* **174:** 265-271.

Bureau, T.E. and Wessler, S.R. 1992. Tourist: a large family of small inverted repeat elements frequently associated with maize genes. *Plant Cell* **4:** 1283-1294.

Bureau, T.E. and Wessler, S.R. 1994. Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell* **6:** 907-916.

Calvi, B.R., Hong, T.J., Findley, S.D., and Gelbart, W.M. 1991. Evidence for a common evolutionary origin of inverted repeat transposons in Drosophila and plants: hobo, Activator, and Tam3. *Cell* **66:** 465-471.

Capy, P., Vitalis, R., Langin, T., Higuet, D., and Bazin, C. 1996. Relationships between transposable elements based upon the integrase-transposase domains: is there a common ancestor? *J Mol Evol* **42:** 359-368.

Caspi, A. and Pachter, L. 2006. Identification of transposable elements using multiple alignments of related genomes. *Genome Res* **16:** 260-270.

Chalvet, F., Grimaldi, C., Kaper, F., Langin, T., and Daboussi, M.J. 2003. Hop, an active Mutator-like element in the genome of the fungus Fusarium oxysporum. *Mol Biol Evol* **20:** 1362-1375.

Chandler, V., Rivin, C., and Walbot, V. 1986. Stable non-mutator stocks of maize have sequences homologous to the Mu1 transposable element. *Genetics* **114:** 1007-1021.

Chomet, P., Lisch, D., Hardeman, K.J., Chandler, V.L., and Freeling, M. 1991. Identification of a regulatory transposon that controls the Mutator transposable element system in maize. *Genetics* **129:** 261-270.

Chopra, S., Brendel, V., Zhang, J., Axtell, J.D., and Peterson, T. 1999. Molecular characterization of a mutable pigmentation phenotype and isolation of the first active transposable element from Sorghum bicolor. *Proc Natl Acad Sci U S A* **96:** 15330-15335.

Comelli, P., Konig, J., and Werr, W. 1999. Alternative splicing of two leading exons partitions promoter activity between the coding regions of the maize homeobox gene Zmhox1a and Trap (transposon-associated protein). *Plant Mol Biol* **41:** 615-625.

de Queiroz, M.V. and Daboussi, M.J. 2003. Impala, a transposon from Fusarium oxysporum, is active in the genome of Penicillium griseoroseum. *FEMS Microbiol Lett* **218:** 317-321.

Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O., and Salzberg, S.L. 1999. Alignment of whole genomes. *Nucleic Acids Res* **27:** 2369-2376.

DeMarco, R., Venancio, T.M., and Verjovski-Almeida, S. 2006. SmTRC1, a novel Schistosoma mansoni DNA transposon, discloses new families of animal and fungi transposons belonging to the CACTA superfamily. *BMC Evol Biol* **6:** 89.

Doak, T.G., Doerder, F.P., Jahn, C.L., and Herrick, G. 1994. A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common "D35E" motif. *Proc Natl Acad Sci U S A* **91:** 942-946.

Edgar, R.C. and Myers, E.W. 2005. PILER: identification and classification of genomic repeats. *Bioinformatics* **21 Suppl 1:** i152-158.

Eisen, J.A., Benito, M.I., and Walbot, V. 1994. Sequence similarity of putative transposases links the maize Mutator autonomous element and a group of bacterial insertion sequences. *Nucleic Acids Res* **22:** 2634-2636.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9:** 18.

Emmons, S.W., Yesner, L., Ruan, K.S., and Katzenberg, D. 1983. Evidence for a transposon in Caenorhabditis elegans. *Cell* **32:** 55-65.

Esposito, T., Gianfrancesco, F., Ciccodicola, A., Montanini, L., Mumm, S., D'Urso, M., and Forabosco, A. 1999. A novel pseudoautosomal human gene encodes a putative protein similar to Ac-like transposases. *Hum Mol Genet* **8:** 61-67.

Feldmar, S. and Kunze, R. 1991. The ORFa protein, the putative transposase of maize transposable element Ac, has a basic DNA binding domain. *EMBO J* **10:** 4003-4010.

Feschotte, C., Osterlund, M.T., Peeler, R., and Wessler, S.R. 2005. DNA-binding specificity of rice mariner-like transposases and interactions with Stowaway MITEs. *Nucleic Acids Res* **33:** 2153-2165.

Feschotte, C. and Pritham, E.J. 2007. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* **41:** 331-368.

Feschotte, C., Swamy, L., and Wessler, S.R. 2003. Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). *Genetics* **163:** 747-758.

Frank, M.J., Liu, D., Tsay, Y.F., Ustach, C., and Crawford, N.M. 1997. Tag1 is an autonomous transposable element that shows somatic excision in both Arabidopsis and tobacco. *Plant Cell* **9:** 1745-1756.

Frey, M., Reinecke, J., Grant, S., Saedler, H., and Gierl, A. 1990. Excision of the En/Spm transposable element of Zea mays requires two element-encoded proteins. *EMBO J* **9:** 4037-4044.

Fujino, K., Matsuda, Y., and Sekiguchi, H. 2009. Transcriptional activity of rice autonomous transposable element Dart. *J Plant Physiol* **166:** 1537-1543.

Grzebelus, D. and Simon, P.W. 2009. Diversity of DcMaster-like elements of the PIF/Harbinger superfamily in the carrot genome. *Genetica* **135:** 347-353.

Hancock, C.N., Zhang, F., and Wessler, S.R. Transposition of the Tourist-MITE mPing in yeast: an assay that retains key features of catalysis by the class 2 PIF/Harbinger superfamily. *Mob DNA* **1:** 5.

Hartl, D.L., Lohe, A.R., and Lozovskaya, E.R. 1997. Modern thoughts on an ancyent marinere: function, evolution, regulation. *Annu Rev Genet* **31:** 337-358.

Hehl, R., Nacken, W.K., Krause, A., Saedler, H., and Sommer, H. 1991. Structural analysis of Tam3, a transposable element from Antirrhinum majus, reveals homologies to the Ac element from maize. *Plant Mol Biol* **16:** 369-371.

Henikoff, S. 1992. Detection of Caenorhabditis transposon homologs in diverse organisms. *New Biol* **4:** 382-388.

Hershberger, R.J., Benito, M.I., Hardeman, K.J., Warren, C., Chandler, V.L., and Walbot, V. 1995. Characterization of the major transcripts encoded by the regulatory MuDR transposable element of maize. *Genetics* **140:** 1087-1098.

Hershberger, R.J., Warren, C.A., and Walbot, V. 1991. Mutator activity in maize correlates with the presence and expression of the Mu transposable element Mu9. *Proc Natl Acad Sci U S A* **88:** 10198-10202.

Holligan, D., Zhang, X., Jiang, N., Pritham, E.J., and Wessler, S.R. 2006. The transposable element landscape of the model legume Lotus japonicus. *Genetics* **174:** 2215-2228.

Holyoake, A.J. and Kidwell, M.G. 2003. Vege and Mar: two novel hAT MITE families from Drosophila willistoni. *Mol Biol Evol* **20:** 163-167.

Jacobson, J.W. and Hartl, D.L. 1985. Coupled instability of two X-linked genes in Drosophila mauritiana: germinal and somatic mutability. *Genetics* **111:** 57-65.

Jacobson, J.W., Medhora, M.M., and Hartl, D.L. 1986. Molecular structure of a somatically unstable transposable element in Drosophila. *Proc Natl Acad Sci U S A* **83:** 8684-8688.

Jiang, N., Bao, Z., Zhang, X., Eddy, S.R., and Wessler, S.R. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431:** 569-573.

Jiang, N., Bao, Z., Zhang, X., Hirochika, H., Eddy, S.R., McCouch, S.R., and Wessler, S.R. 2003. An active DNA transposon family in rice. *Nature* **421:** 163-167.

Jurka, J. 2000. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* **16:** 418-420.

Kapitonov, V.V. and Jurka, J. 1999. Molecular paleontology of transposable elements from Arabidopsis thaliana. *Genetica* **107:** 27-37.

Kapitonov, V.V. and Jurka, J. 2004. Harbinger transposons and an ancient HARBI1 gene derived from a transposase. *DNA Cell Biol* **23:** 311-324.

Kempken, F. and Kuck, U. 1996. restless, an active Ac-like transposon from the fungus Tolypocladium inflatum: structure, expression, and alternative RNA splicing. *Mol Cell Biol* **16:** 6563-6572.

Koga, A., Suzuki, M., Inagaki, H., Bessho, Y., and Hori, H. 1996. Transposable element in fish. *Nature* **383:** 30.

Kuang, H., Padmanabhan, C., Li, F., Kamei, A., Bhaskar, P.B., Ouyang, S., Jiang, J., Buell, C.R., and Baker, B. 2009. Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: new functional implications for MITEs. *Genome Res* **19:** 42-56.

Kurtz, S. and Schleiermacher, C. 1999. REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics* **15:** 426-427.

Lampe, D.J., Walden, K.K., and Robertson, H.M. 2001. Loss of transposase-DNA interaction may underlie the divergence of mariner family transposable elements and the ability of more than one mariner to occupy the same genome. *Mol Biol Evol* **18:** 954-961.

Langdon, T., Jenkins, G., Hasterok, R., Jones, R.N., and King, I.P. 2003. A high-copy-number CACTA family transposon in temperate grasses and cereals. *Genetics* **163:** 1097-1108.

Lisch, D. 2002. Mutator transposons. *Trends Plant Sci* **7:** 498-504.

Lisch, D., Girard, L., Donlin, M., and Freeling, M. 1999. Functional analysis of deletion derivatives of the maize transposon MuDR delineates roles for the MURA and MURB proteins. *Genetics* **151:** 331-341.

Lisch, D.R., Freeling, M., Langham, R.J., and Choy, M.Y. 2001. Mutator transposase is widespread in the grasses. *Plant Physiol* **125:** 1293-1303.

Makarova, K.S., Aravind, L., and Koonin, E.V. 2002. SWIM, a novel Zn-chelating domain present in bacteria, archaea and eukaryotes. *Trends Biochem Sci* **27:** 384-386.

Masson, P., Strem, M., and Fedoroff, N. 1991. The tnpA and tnpD gene products of the Spm element are required for transposition in tobacco. *Plant Cell* **3:** 73-85.

McCarthy, E.M. and McDonald, J.F. 2003. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* **19:** 362-367.

McClintock, B. 1947. Cytogenetic studies of maize and *Neurospora*. *Carnegie Inst Washington Year Book* **46:** 146-152.

McClintock, B. 1948. Mutable loci in maize. *Carnegie Inst Washington Year Book* **47:** 155–169.

McClintock, B. 1954. Mutations in maize and chromosomal aberrations in Neurospora. *Carnegie Inst Washington Year Book* **53:** 254-260.

McClure, M.A., Richardson, H.S., Clinton, R.A., Hepp, C.M., Crowther, B.A., and Donaldson, E.F. 2005. Automated characterization of potentially active retroid agents in the human genome. *Genomics* **85:** 512-523.

Moreno-Vazquez, S., Ning, J., and Meyers, B.C. 2005. hATpin, a family of MITE-like hAT mobile elements conserved in diverse plant species that forms highly stable secondary structures. *Plant Mol Biol* **58:** 869-886.

Naito, K., Cho, E., Yang, G., Campbell, M.A., Yano, K., Okumoto, Y., Tanisaka, T., and Wessler, S.R. 2006. Dramatic amplification of a rice transposable element during recent domestication. *Proc Natl Acad Sci U S A* **103:** 17620-17625.

Neuveglise, C., Chalvet, F., Wincker, P., Gaillardin, C., and Casaregola, S. 2005. Mutator-like element in the yeast Yarrowia lipolytica displays multiple alternative splicings. *Eukaryot Cell* **4:** 615-624.

Pearson, W.R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol* **183:** 63-98.

Peterson, P.A. 1953. A mutable pale green locus in maize *Genetics* **38:** 682-683.

Pevzner, P.A., Tang, H., and Tesler, G. 2004. De novo repeat classification and fragment assembly. *Genome Res* **14:** 1786-1796.

Pietrokovski, S. and Henikoff, S. 1997. A helix-turn-helix DNA-binding motif predicted for transposases of DNA transposons. *Mol Gen Genet* **254:** 689-695.

Plasterk, R.H. 1996. The Tc1/mariner transposon family. *Curr Top Microbiol Immunol* **204:** 125-143.

Price, A.L., Jones, N.C., and Pevzner, P.A. 2005. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1:** i351-358.

Qin, M.M., Robertson, D.S., and Ellingboe, A.H. 1991. Cloning of the Mutator transposable element MuA2, a putative regulator of somatic mutability of the a1-Mum2 allele in maize. *Genetics* **129:** 845-854.

Ray, D.A., Pagan, H.J., Thompson, M.L., and Stevens, R.D. 2007. Bats with hATs: evidence for recent DNA transposon activity in genus Myotis. *Mol Biol Evol* **24:** 632-639.

Robertson, D.S. 1978. Characterization of a mutator system in maize. *Mutation Research* **Volume 51:** Pages 21-28

Tian, P.F. 2006. Progress in plant CACTA elements. *Yi Chuan Xue Bao* **33:** 765-774.

Tu, Z. 2001. Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, Anopheles gambiae. *Proc Natl Acad Sci U S A* **98:** 1699-1704.

Tudor, M., Lobocka, M., Goodell, M., Pettitt, J., and O'Hare, K. 1992. The pogo transposable element family of Drosophila melanogaster. *Mol Gen Genet* **232:** 126-134.

Volfovsky, N., Haas, B.J., and Salzberg, S.L. 2001. A clustering method for repeat analysis in DNA sequences. *Genome Biol* **2:** RESEARCH0027.

Wang, Q. and Dooner, H.K. 2006. Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. *Proc Natl Acad Sci U S A* **103:** 17644-17649.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O. et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8:** 973-982.

Xu, Z. and Wang, H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res* **35:** W265-268.

Xu, Z., Yan, X., Maurais, S., Fu, H., O'Brien, D.G., Mottinger, J., and Dooner, H.K. 2004. Jittery, a Mutator distant relative with a paradoxical mobile behavior: excision without reinsertion. *Plant Cell* **16:** 1105-1114.

Yang, G., Nagel, D.H., Feschotte, C., Hancock, C.N., and Wessler, S.R. 2009. Tuned for transposition: molecular determinants underlying the hyperactivity of a Stowaway MITE. *Science* **325:** 1391-1394.

Yang, G., Zhang, F., Hancock, C.N., and Wessler, S.R. 2007. Transposition of the rice miniature inverted repeat transposable element mPing in Arabidopsis thaliana. *Proc Natl Acad Sci U S A* **104:** 10962-10967.

Yu, Z., Wright, S.I., and Bureau, T.E. 2000. Mutator-like elements in Arabidopsis thaliana. Structure, diversity and evolution. *Genetics* **156:** 2019-2031.

Zabala, G. and Vodkin, L. 2007. Novel exon combinations generated by alternative splicing of gene fragments mobilized by a CACTA transposon in Glycine max. *BMC Plant Biol* **7:** 38.

Zabala, G. and Vodkin, L.O. 2005. The wp mutation of Glycine max carries a gene-fragment-rich transposon of the CACTA superfamily. *Plant Cell* **17:** 2619-2632.

Zhang, L., Dawson, A., and Finnegan, D.J. 2001a. DNA-binding activity and subunit interaction of the mariner transposase. *Nucleic Acids Res* **29:** 3566-3575.

Zhang, X., Feschotte, C., Zhang, Q., Jiang, N., Eggleston, W.B., and Wessler, S.R. 2001b. P instability factor: an active maize transposon system associated with the amplification of Tourist-like MITEs and a new superfamily of transposases. *Proc Natl Acad Sci U S A* **98:** 12572-12577.

Zhang, X., Jiang, N., Feschotte, C., and Wessler, S.R. 2004. PIF- and Pong-like transposable elements: distribution, evolution and relationship with Tourist-like miniature inverted-repeat transposable elements. *Genetics* **166:** 971-986.

Chapter 2

TARGeT -- A web based pipeline for retrieving and characterizing gene and transposable element families from genomic sequences

---

Yujun Han, James M. Burnette III and Susan R. Wessler.

Nucleic Acids Research. Vol. 37, No. 11. 2009.

Abstract

Gene families compose a large proportion of eukaryotic genomes. The rapidly-expanding genomic sequence database provides a good opportunity to study gene family evolution and function. However, most gene family identification programs are restricted to searching protein databases where data is often lagging behind the genomic sequence data. Here I report a user friendly web based pipeline, named TARGeT (Tree Analysis of Related Genes and Transposons), which uses either a DNA or amino acid “seed” query to: (1) automatically identify and retrieve gene family homologs from a genomic database, (2) characterize gene structure, and (3) perform phylogenetic analysis. Due to its high speed, TARGeT is also able to characterize very large gene families, including transposable elements. I evaluated TARGeT using well annotated datasets, including the ascorbate peroxidase gene family of rice, maize and sorghum, and several transposable element families in rice. In all cases TARGeT rapidly recapitulated the known homologs and predicted new ones. I also demonstrated that TARGeT outperforms similar pipelines and has functionality that is not offered elsewhere.

Introduction

A major discovery of eukaryote genome projects is that unexpectedly large numbers of genes are members of gene families. Gene families comprise 49% of the genes in *Caenorhabditis elegans*, 41% in *Drosophila melanogaster*, 38% in *Homo sapiens*, 65% in *Arabidopsis thaliana* and 77% in *Oryza sativa* L. ssp. Japonica (Goff *et al.* 2002; Initiative 2000a; Li *et al.* 2001; Rubin *et al.* 2000; Yu *et al.* 2002). That the size of gene families often varies among closely related species indicates that gene duplication and gene family diversification is an ongoing process (Lespinet *et al.* 2002; Lyckegaard and Clark 1989; Neitz and Neitz 1995).

Duplicate genes arise in several ways including whole-genome duplication (Dehal and Boore 2005; Wendel 2000; Wolfe and Shields 1997) and segmental duplication (Cheung *et al.* 2003; Eichler 2001). Segmental duplication events can be further classified into tandem and interspersed (Hurles 2004). A tandem duplication event can result from either homologous (Bailey *et al.* 2003) or nonhomologous recombination mechanisms (Koszul *et al.* 2004), while interspersed duplication events are mainly caused by the activity of transposable elements (Jiang *et al.* 2004; Morgante *et al.* 2005; Tchenio *et al.* 1993; Vanin 1985).

Gene family members can be detected by clustering genes based on their similarity (Dayhoff 1976; Heger and Holm 2000), and new members can be identified through similarity comparison to known members. Many gene family databases have been established, including Pfam (Finn *et al.* 2008), TreeFam (Li *et al.* 2006) and PANTHER (Mi *et al.* 2005), etc. While these gene family databases are useful

recourses, they are not updated at the same rapid pace as that of newly generated genomic sequences. Researchers interested in particular gene families often have to perform their own searches to obtain the most current collection of sequences.

The identification of gene family members using sequence similarity searches is often complicated by the detection of homologs from other gene families. Phylogenetic analysis is a powerful tool to identify homologs of interest and to provide additional information about gene function and evolution. To this end, researchers can perform manual searches using publicly available programs such as BLAT (Kent 2002), Wise2 (Birney *et al.* 2004), BLAST (Altschul *et al.* 1997), FASTA (Pearson 1990) and HMMER (Eddy 1998), followed by sequence alignment and phylogenetic analysis. However, these procedures can be complicated as they often require extensive manual curation, particularly if homologous regions need to be extracted from genomic sequences. While this is a manageable problem for a small gene family, it can be a tedious and time-consuming process when the target gene family is large. More significantly, the quality of the results often suffers.

In addition to the more traditional gene families, transposable elements (TEs) can also be viewed as members of “special” gene families that are able to duplicate themselves by the activity of element-encoded proteins. TEs often constitute the largest component of eukaryotic genomes and their identification and classification are essential to accurate genome annotation (Lander *et al.* 2001; Meyers *et al.* 2001). However, as with large gene families, the very high copy numbers of some TEs makes their retrieval from genomic sequence and characterization an extremely difficult task.

The increasing pace of genomic sequencing projects demands a computer-assisted pipeline that can rapidly and accurately identify and characterize gene families.

Several automated pipelines have been developed to ease homolog identification and most of them are limited to protein or EST databases. For example, PhyloBLAST (Brinkman *et al.* 2001), Pyphy (Sicheritz-Ponten and Andersson 2001), HoSeql (Arigon *et al.* 2006), PhyloGena (Hanekamp *et al.* 2007), and TRIBE-MCL (Enright *et al.* 2002) perform BLASTP searches and retrieve data from protein databases. SimESTs uses TBLASTN to search EST databases (Frank *et al.* 2006). Because these programs only compare protein coding sequences, they will miss any mutational events that occur within noncoding regions.

TARGeT (Tree Analysis of Related Genes and Transposons) is a program to streamline the process of retrieving, annotating and analyzing both gene families and TE families from a genomic database. The core of the TARGeT pipeline is an algorithm called Putative Homolog Identifier (PHI) that uses a series of steps to predict gene structure using BLAST results. From the predicted gene structure, PHI extracts the amino acid sequences of putative homologs for use in subsequent phylogenetic analysis. I have compared TARGeT with two pipelines, FGF and GFScan, which can also be used to retrieve gene families from genomic databases. Results are presented showing that TARGeT significantly outperforms both programs and adds several layers of functionality not present in existing programs. To make it easier for users, especially non-specialists, TARGeT was implemented as a user-friendly web-based pipeline (http://target.iplantcollaborative.org/). All initial input for TARGeT is organized on a web

form and the results are presented in the browser. All results and supporting files are documented and are available for download. TARGeT provides several points where results can be inspected and analyses can be repeated.

Methods

TARGeT can use either protein or DNA sequences as the query. BLASTN searches are used for DNA queries while TBLASTN is used for protein queries. The pipeline that uses TBLASTN is the focus of this paper because it is more complex and may have wider application. TARGeT uses Muscle (Edgar 2004) to calculate the multiple alignment and TreeBest (Li *et al.* 2006) to generate the phylogenetic tree of the putative homologs with the neighbour joining method (Saitou and Nei 1987). The other functions of TARGeT are carried out by several Perl scripts developed by the authors.

Rice genomic data was obtained from Genbank (Burks *et al.* 1985; Karsch-Mizrachi and Ouellette 2001) with accession numbers from NC_008394 to NC_008405. Maize genomic data was downloaded from the Maize Genome Sequencing Project (http://www.maizesequence.org; version: Dec. 2008). Sorghum genomic data was from the Sorghum bicolor genome project (http://www.jgi.doe.gov; version: 2008 Sorbi1 assembly).

There are five main steps in the TARGeT pipeline with a checkpoint at the end of each: (A) preparation of the query when multiple sequences are to be submitted, (B) BLAST search (either BLASTN or TBLASTN), (C) homolog prediction, (D) multiple alignment and (E) phylogenetic tree estimation (**Figure 2.1**). Details of each step are presented in the results section using the ascorbate peroxidase (APx) gene family as an example.

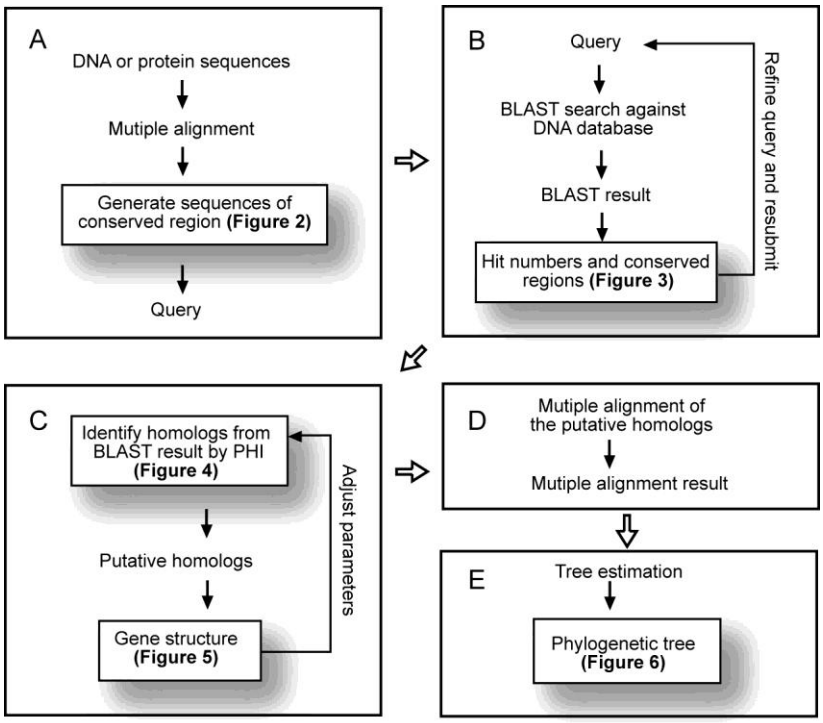


Figure 2.1 Map of the five main steps of the TARGeT pipeline. Users are able to inspect the results of each step before going on to the next step. (A) Preparation of the query when more than one sequence is being used. This is an optional step and its output is shown in **Figure 2.2**. (B) BLAST search. Results are shown in Figure 3. (C) Homolog identification by PHI. The algorithm is explained in **Figure 2.4** and the result of this step is shown in **Figure 2.5**. (D) Multiple alignment. (E) Tree building. A phylogenetic tree is shown in **Figure 2.6**.

TARGeT can be accessed on a web server, where all data used and generated by TARGeT is entered in a log file. TARGeT output is presented in a single webpage that uses nested tabs to organize the data, images, and re-submission forms for each TARGeT run during a session. There is a final tab for each run called Provenance where the user can view the parameters used by TARGeT in a log file and also download an archive that includes all files and images for offline viewing and analysis. The output includes the XML log file, BLAST results in image and text format, PHI results in image and text format, multiple alignment in FASTA format, and the phylogenetic tree in Newick and jpeg formats.

Results

1 Searching for APx gene family in rice. Rice and *Arabidopsis* serve as model plant monocot and dicot species, respectively. They diverged from a common ancestor about 200 million years ago (Yang *et al.* 1999) and their genomes are fully sequenced (Goff *et al.* 2002; Initiative 2000b; Yu *et al.* 2002). Thus they provide excellent opportunities to evaluate the cross-species searching ability of TARGeT. I searched the rice APx gene family using the *Arabidopsis* APx protein sequences as query and compared the results generated by TARGeT to the published data. The goal of this exercise was to see how well TARGeT would perform at predicting the rice APx family members. I chose APx because it is a small but important gene family that has been well annotated in both *Arabidopsis* and rice. Based on the literature, there are as many as 9 APx family members in *Arabidopsis* (Mittler et al. 2004) and 8 in rice (Teixeira et al. 2006) (**Table 2.1**). The APx family shares sequence similarity with several other peroxidase families (Passardi *et al.* 2007) and, as such, is a good dataset to test the ability of TARGeT to discriminate between closely related protein families.

Table 2.1 The APx gene family homologs of *Arabidopsis*, rice, maize and sorghum.

Rice APX Family	TATE's ID	Matched		Missing		Wrong	
		bps	%	bps	%	bps	%
OsAPX1	T_APx_5	712	97.53	18	2.47	6	0.82
OsAPX2	T_APx_10	724	97.71	17	2.29	5	0.67
OsAPX3	T_APx_7	721	97.7	17	2.3	5	0.68
OsAPX4	T_APx_2	727	98.51	11	1.49	8	1.08
OsAPX5	T_APx_6	778	98.61	11	1.39	11	1.39
OsAPX6	T_APx_4	778	94.99	41	5.01	8	0.98
OsAPX7	T_APx_3	780	97.01	24	2.99	6	0.75
OsAPX8	T_APx_8	774	98.47	12	1.53	12	1.53
Average		749.25	97.57	18.88	2.43	7.63	0.99

1.1 BLAST Search. To improve the chances of finding target gene family members, multiple queries can be submitted as long as they are homologs. An optional multiple alignment step is provided for users to select sequences from conserved regions (**Figure 2.1A**). As an example, for the APx gene family I selected as query the sequences from the well-aligned (boxed) region in **Figure 2.2**.

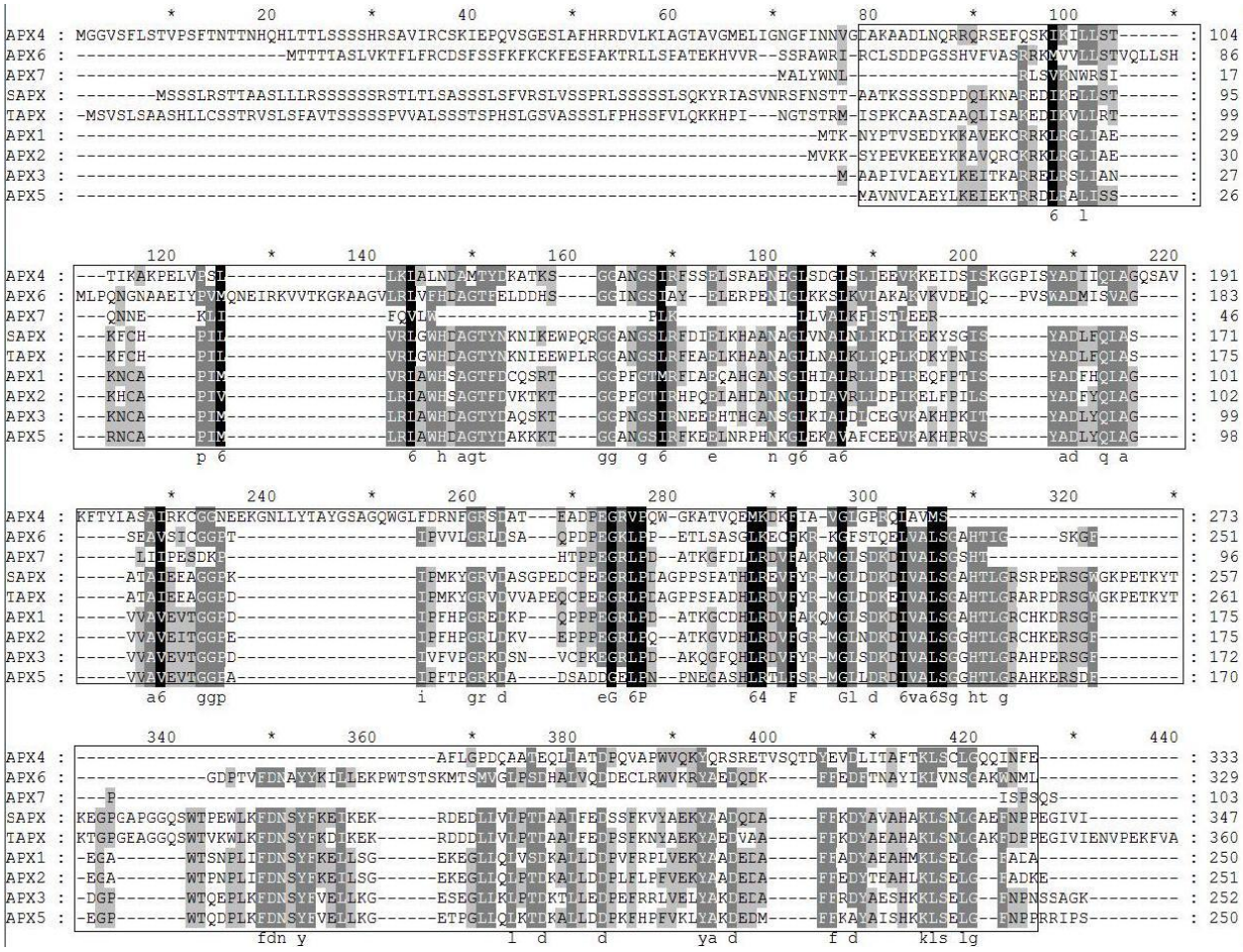


Figure 2.2 Multiple alignment of *Arabidopsis* APx protein sequences. Sequences in the boxed region were extracted to form the query sequences. APx7 was not included because it aligns poorly.

To aid users in viewing the BLAST result, TARGeT produces an image showing a rough estimation of BLAST High Scoring Pair (HSP) numbers and conserved regions along the length of each query sequence (**Figure 2.1B and Figure 2.3**). This is helpful for a quick overall view especially when the BLAST output is large. In this way the user

can see the information used by TARGeT and, if necessary, modify the query in a subsequent BLAST search. For example, TAPX, which is one of the *Arabidopsis* APx genes, is 426 amino acids. With the full length sequence as the query, low copy regions can be detected at the beginning and at the end (**Figure 2.3A**). Readers should note that the number of HSPs (up to 50) is much larger than the number of known APx genes in the rice genome. This inconsistency is largely due to the existence of other gene families that share sequence similarity with the APx gene family. As shown in later steps, true homologs belonging to the APx gene family will be discerned from those of other families. Using the full length TAPX sequence as the query, only three APx homologs were found in rice (data not shown). However, 5 APx homologs were found when the sequence from the boxed region was used as the new query (**Figure 2.3B**).

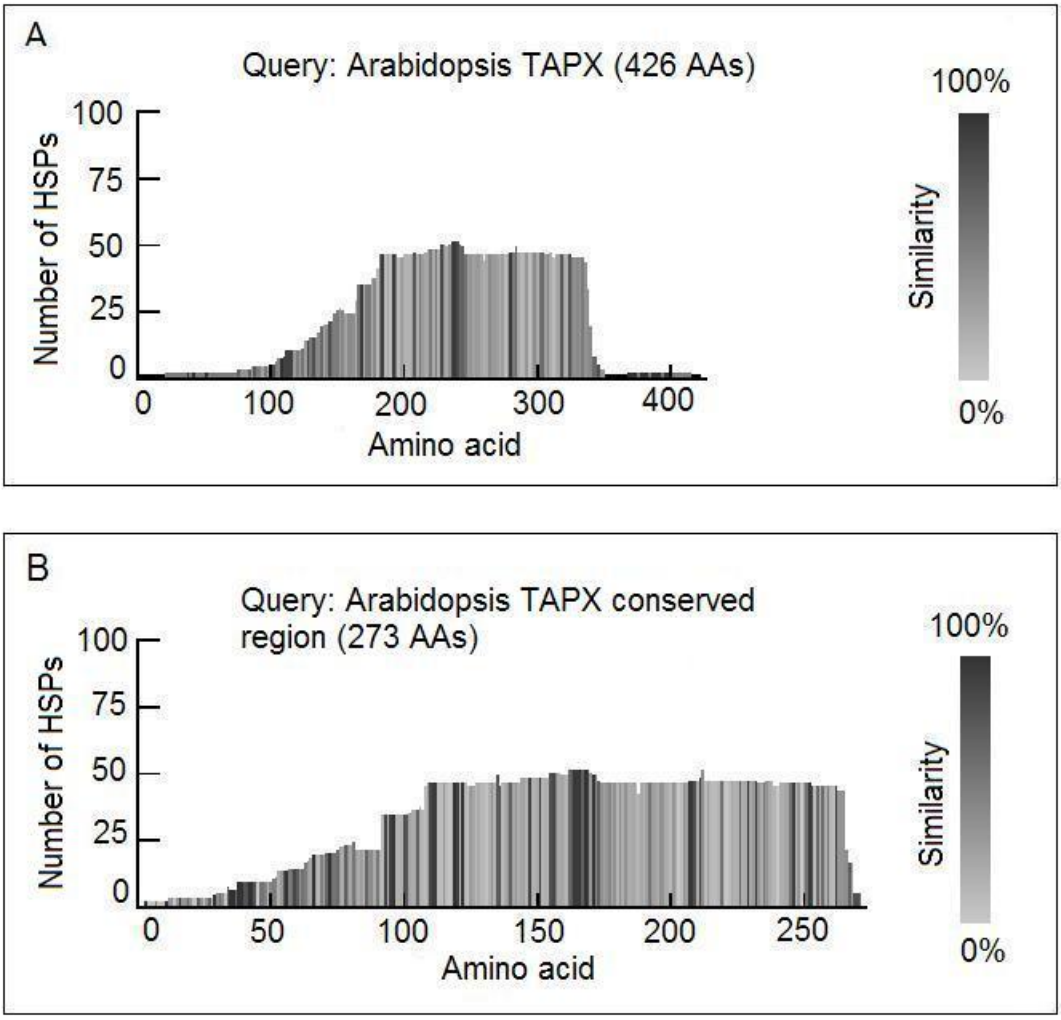
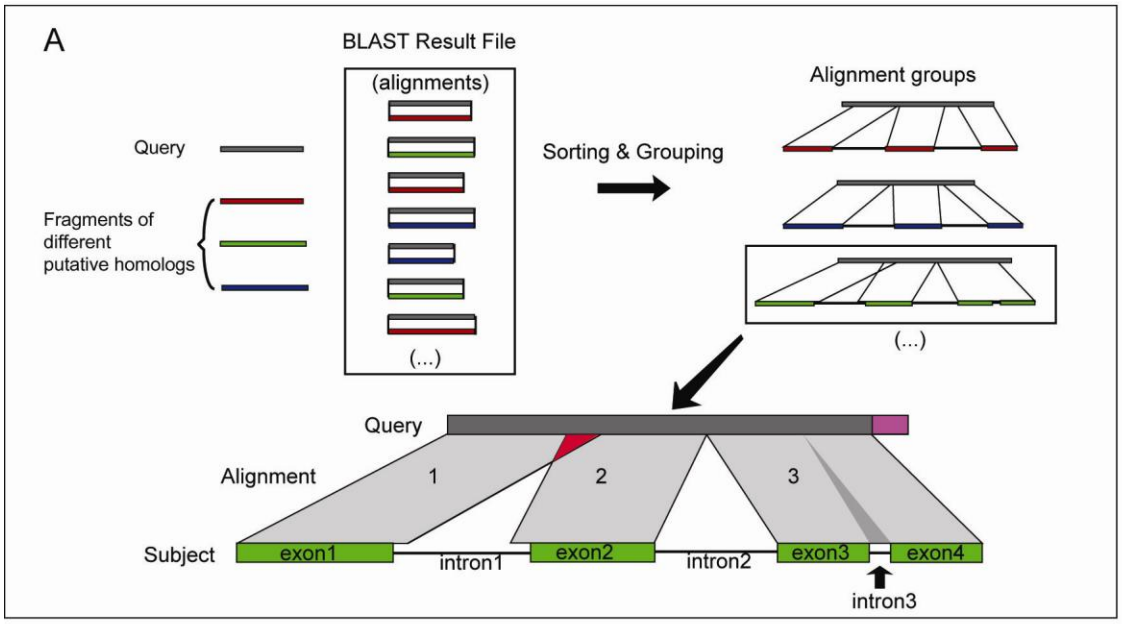
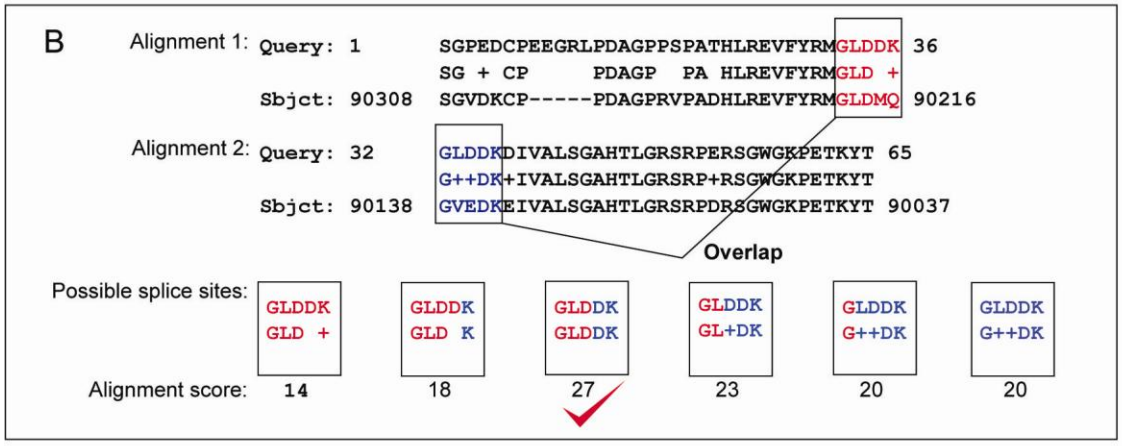


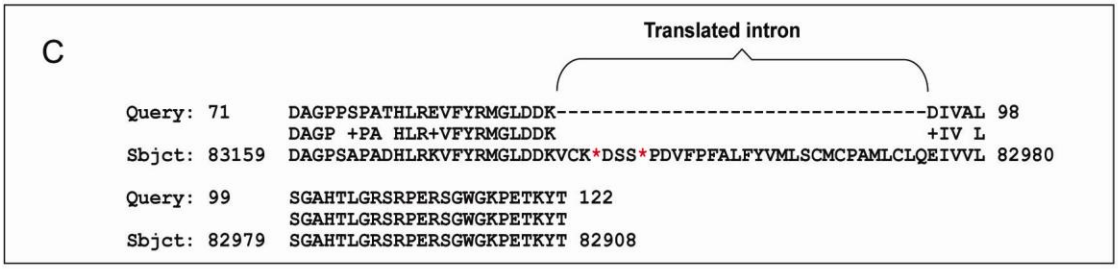
Figure 2.3 TARGeT output provides a rough visualization of the BLAST result. X-axis is the length of the query; Y-axis is the number of BLAST HSPs. The gray-gradient shows the similarity which is calculated by dividing the sum of identities and similarities by the number of the aligned amino acids along the HSP. Darker represents higher similarity at that position.

1.2 Putative Homolog Identification. Several factors make it difficult to identify reliable homologs from BLAST output and result in a high false positive rate (Frickey and Lupas 2004; Koski and Golding 2001; Xuan *et al.* 2002). Lack of explicit treatment of frameshifts and introns is also a disadvantage of TBLASTN (Gertz *et al.* 2006). To solve these problems, I developed a program called **PHI** (Putative Homolog Identifier), which takes into account the e-value (default 0.01) as well as a second parameter called the minimal match percentage (**MMP, defaults to 70%)** to find reliable homologs. The two main stages in PHI (grouping and refinement) are explained below.

1.2.1 Grouping. Introns or low similarity regions can break a complete alignment into smaller HSPs. In addition when a frameshift occurs, TBLASTN produces separate HSPs. To retrieve the intact sequence of each homolog or pseudogene, PHI sorts the HSPs based on position and strand in the genomic sequence. In this step, HSPs that are from the same homolog are grouped together by the sequence position of query and subject (**Figure 2.4A,** top part)**.** Two HSPs are assumed to belong to different groups if they are separated by a distance greater than the minimum intron length (a parameter adjustable by the user, defaults to 8,000 nt) or if they are on different strands. When there is more than one way to connect the HSPs (which can happen when there are repetitive domains in the query), PHI uses an overall HSP score to determine the correct order. A match percentage is calculated by dividing the sum length of the matches in each group by the length of the query. If this number is greater than the **MMP**, then the group is sent to the refinement stage. HSPs that fail to satisfy the MMP are available to interested users as a record file.


A
BLAST Result File
(alignments)
(...)
Query
Fragments of different putative homologs
Sorting & Grouping
Alignment groups
(...)
Query
Alignment
1
2
3
Subject
exon1
intron1
exon2
intron2
exon3
exon4
intron3


B
Alignment 1: Query: 1 SGPEDCPEEGRLPDAGPPSPATHLREVFYRMGLDDK 36
SG + CP PDAGP PA HLREVFYRMGLD +
Sbjct: 90308 SGVDKCP-----PDAGPRVPADHLREVFYRMGLDMQ 90216
Alignment 2: Query: 32 GLDDKDIVALSGAHTLGRSRPERSGWGKPETKYT 65
G++DK+IVALSGAHTLGRSRP+RSGWGKPETKYT
Sbjct: 90138 GVEDKEIVALSGAHTLGRSRPDRSGWGKPETKYT 90037
Overlap
Possible splice sites:
GLDDK GLD +
GLDDK GLD K
GLDDK GLDDK
GLDDK GL+DK
GLDDK G++DK
GLDDK G++DK
Alignment score:
14
18
27
23
20
20


C
Translated intron
Query: 71 DAGPPSPATHLREVFYRMGLDDK----------DIVAL 98
DAGP +PA HLR+VFYRMGLDDK +IV L
Sbjct: 83159 DAGPSAPADHLRKVFYRMGLDDKVCK*DSS*PDVFPFALFYVMLSCMCPAMLCLQEIVVL 82980
Query: 99 SGAHTLGRSRPERSGWGKPETKYT 122
SGAHTLGRSRPERSGWGKPETKYT
Sbjct: 82979 SGAHTLGRSRPERSGWGKPETKYT 82908

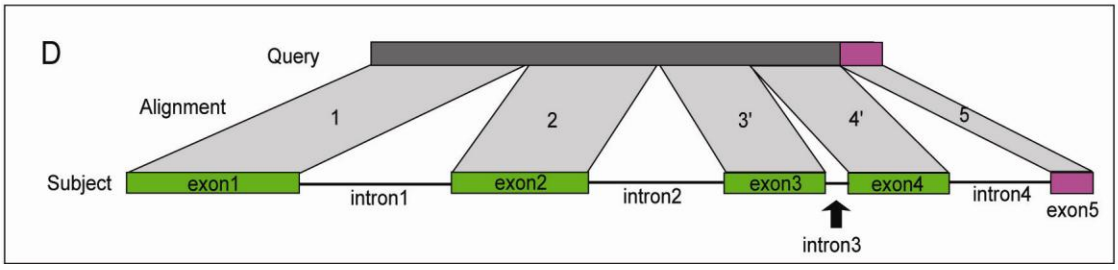

D
Query
Alignment
1
2
3'
4'
5
Subject
exon1
intron1
exon2
intron2
exon3
exon4
intron4
exon5
intron3

Figure 2.4 The sorting and refinement stages of the Putative Homolog Identifier (PHI) program. See the text for details. (A) In the grouping stage, alignments are sorted and grouped. Dark bars are queries and colored bars are homologs. Each group corresponds to one putative homolog. The green group is shown in detail to illustrate potential problems. (B) Two overlapping HSPs together with 6 possible alternative positions are shown. The separation that produces the highest score in the overlapping region is noted with a red check. (C) An HSP that includes an intron. The intron is detected and cut out by PHI, resulting in two separated HSPs. Red asterisks represent premature stop codons. (D) Figure presentation of the result after the refinement stage. There is no overlap between HSPs 1 and 2. HSP 3 in (C) is separated by the small intron into new HSP 3' and 4'. An additional exon (5) was found and is shown in pink.

1.2.2 Refinement . After the grouping step, several potential problems often remain in the HSPs of each group. A demonstration figure to illustrate some problems is shown in the lower part of **Figure 2.4A**. First, there is an overlap (indicated by a red triangle) between HSP 1 and HSP 2; second, there is an intron (darker area) that has been falsely translated and included in HSP 3; finally, there is a small area in the query (pink region) that has no HSP in the BLAST result, which results in the failure to detect a small exon due to its insignificant e-value. In the refinement stage, the most likely split position is detected within the overlapping region. Introns in the HSPs are removed, and a second round BLAST search is performed to find the missing exons. Several result

files will be generated after this step, including the homologous sequences in both DNA and protein FASTA formats.

Resolving the boundary between two overlapping HSPs. In TBLASTN outputs, two successive HSPs often overlap due to coincident similarity beyond the true boundaries, resulting in misalignment between the query and the subject. An example is shown in **Figure 2.4B** (boxed regions). In this example, the end of HSP 1 overlaps the beginning of HSP 2 by five amino acids corresponding to amino acids 32–36 of the query (red residues GLDDK), 90212–90216 (red residues GLDMQ) and 90138–90142 (blue residues GVEDK) of the subject. PHI determines the most likely correct boundary by choosing the alignment that has the highest alignment score from all of the possible alignments within the overlapping region. For the example shown in **Figure 2.4B** there are 6 possible alignments. A score is calculated for each alignment using the BLOSUM62 matrix and any amino acid that aligns to a gap or a stop codon will be penalized 12 points. The third alignment in **Figure 2.4B** has the highest score and thus PHI assumes the true boundary in the subject is between the two aspartic acid residues. After the true boundary is located, additional amino acids will be trimmed off of the HSPs (MQ in HSP 1 and GVE in HSP 2). For the rice APx gene family, this step trimmed 21 amino acid residues on average from each homolog.

Identifying small introns. The function of this step is to identify and remove introns that appear as gaps within the HSPs. Any gap in the subject that has a length greater than the minimum intron length parameter (user adjustable parameter, default 60 nt) is identified as an intron and will be removed resulting in two (smaller) new HSPs

(**Figure 2.4C**, **Figure 2.4D**). For each rice APx homolog, TARGeT identified, on average, 1.3 introns corresponding to 41.9 falsely translated amino acids.

Identifying small exons. Small exons will be missed by BLAST searches when their alignments do not meet the e-value cut-off. Such small exons may be found by increasing the e-value. However, for a large database, simply increasing the e-value could increase the computational burden of TARGeT, and there is no guarantee that all exons will be identified because the suitable e-value is unknown. To improve the prediction of small exons, PHI can perform a second round BLAST search, using a small database containing only the sequences of putative homologs (including the predicted intronic and flanking regions). Because e-value calculation is dependent in part on the size of the database, short alignments to the original query sequence(s) may now be significant (**Figure 2.4D)**. For each rice APx homolog, this second round of BLAST identified, on average, 1.6 additional exons and 33.4 amino acids.

1.3 Illustration of PHI Output. After the refinement stage, an image is generated that provides a view of the predicted gene structure for each putative homolog (**Figure 2.1C** and **Supplemental Figure 1**). Features of this image include the similarity between each putative homolog and its query, the locations of exons, introns, premature stop codons (represented by asterisks in the BLAST output), and frameshifts. Frameshifts are identified by comparing HSPs that are close to each other (less than 5 amino acids by default) and are on the same strand but are in different reading frames. In the demonstration figure, putative pseudogenes may be genes with premature stop

codons or frameshifts that are marked with red or blue dots, respectively (**Supplemental Figure 7 and 8).**

Using default parameters, 46 putative rice APx homologs were identified and clustered into two groups based on their gene structures (**Figure 2.5 and Supplemental Figure 1**). There are 11 homologs in the small group, among which, TOAPx_2-8 and TOAPx_10 were found to correspond to known rice APx genes OsAPx1-OsAPx8 (**Table 2.1**). For the remaining 35 putative homologs, comparison of their sequences and gene structures revealed that they are not APx homologs but are instead from other peroxidase gene families (data not shown).

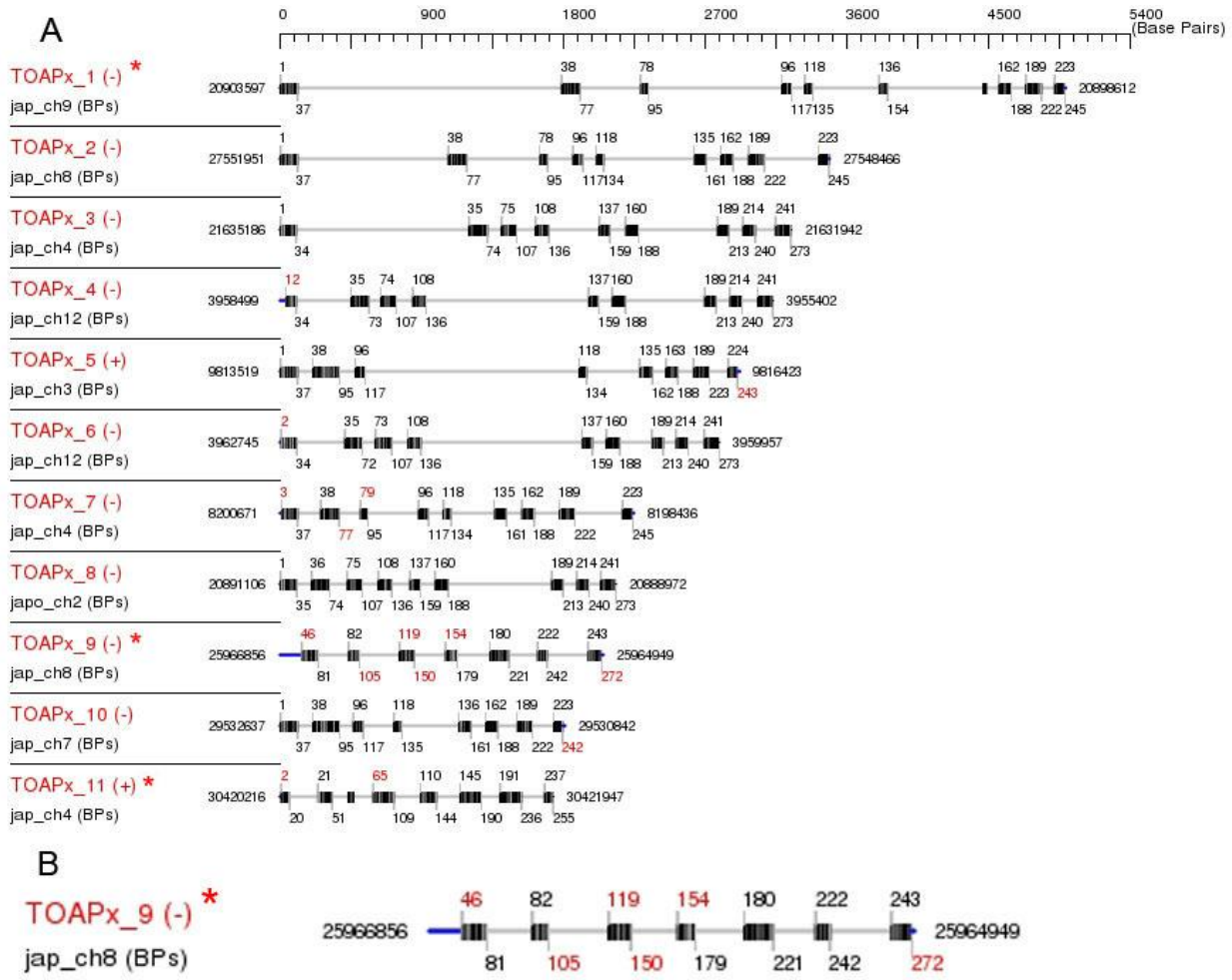


Figure 2.5 TARGeT output of the gene structure of rice APx family members. The small exon searching option was on. The other parameters are the default settings. (A) Exon intron structure of eleven reliable rice APx homologs detected by TARGeT. All 46 putative homologs are in **supplemental Figure 1**. (B) A larger figure of TOAPx_9 from (A). Query and subject names are shown on the left. “+” or “-” indicates the strand of the hit. Unmatched query regions at the ends of each homolog are in blue. Black or gray gradient bars represent the exons. Darker represents higher similarity. Numbers

flanking each gene structure are positions of the subject while numbers above and below the exons are the positions of the query. Red numbers indicate discontinuous predicated exons.

To assess the accuracy of homolog sequences retrieved by TARGeT, I considered two situations (this is not a step of TARGeT). One situation might occur at the ends of the query-target alignment where the program failed to identify some amino acids at the end. I refer to this as “missing” and can occur when the end of homolog sequences are not as well conserved as the sequences within. By comparing the homolog sequence to the query sequence, the numbers of “missing” amino acids were counted manually. For example, if the query is 100 amino acids and the alignment is from 5 to 97, the missing number of this homolog is 4+3 = 7. The “missed” rate is calculated by dividing the number of missed amino acids by the length of the query (7% in the above example). In contrast, I refer to an “error” as a situation when the program incorrectly predicts amino acids within a homolog sequence. By comparing the homolog sequence to the previously published rice APx protein sequence, mismatched amino acids were counted manually as the “error” number of this homolog. The “error” rate is calculated by dividing the number of incorrect amino acid assignments by the length of the corresponding region in the previously published rice APx protein sequence. The missed and error rates may vary for each predicated homolog sequence, because they depend on the level of conservation between the homolog and the query sequences.

For the rice APx example above, the average missed rate is 1.11% and the average error rate is 0.49% (**Table 2.1**).

1.4 Multiple Alignment and Tree Estimation. If users are satisfied with the putative homologs found by TARGeT, they can either download the sequences in FASTA format or let TARGeT use the data to generate a phylogenetic tree. Users also have the option to employ other tree estimation methods by downloading the alignment and using the software of their choice. The phylogenetic tree and the figure showing the tree are generated by TreeBest. When there are many homologs, names on the figure will be difficult to read because the figure size cannot be varied. To solve this problem users can download the newick file and draw the tree by themselves using software such as treeview (Page 1996). I have also provided two more solutions on the server. The first is to use Jalview and the second is to copy the newick format tree file and submit it to PhyloWidget (Jordan and Piel 2008), which is a powerful web based tree viewer.

From the TARGeT-generated tree of APx homologs (**shaded region in Figure 2.6**) it is clear that the known APx family homologs are separated from the other putative homologs. Consideration of both gene structures (**Figure 2.5 and Supplemental Figure 1**) and positions in the phylogenetic tree (**Figure 2.6**) led to the identification of three putative new rice APx genes (TOAPx_1, TOAPx_9 and TOAPx_11) that have high similarity to *Arabidopsis* APx3 (Identities = 80%, Positives = 92%), APx6 (Identities = 62%, Positives = 77%) and APx4 (Identities = 71%, Positives = 82%), respectively. To provide evidence that these are real genes, they were used as

queries against the rice cDNA database in Genbank. Each gene matched several cDNAs (data not shown).

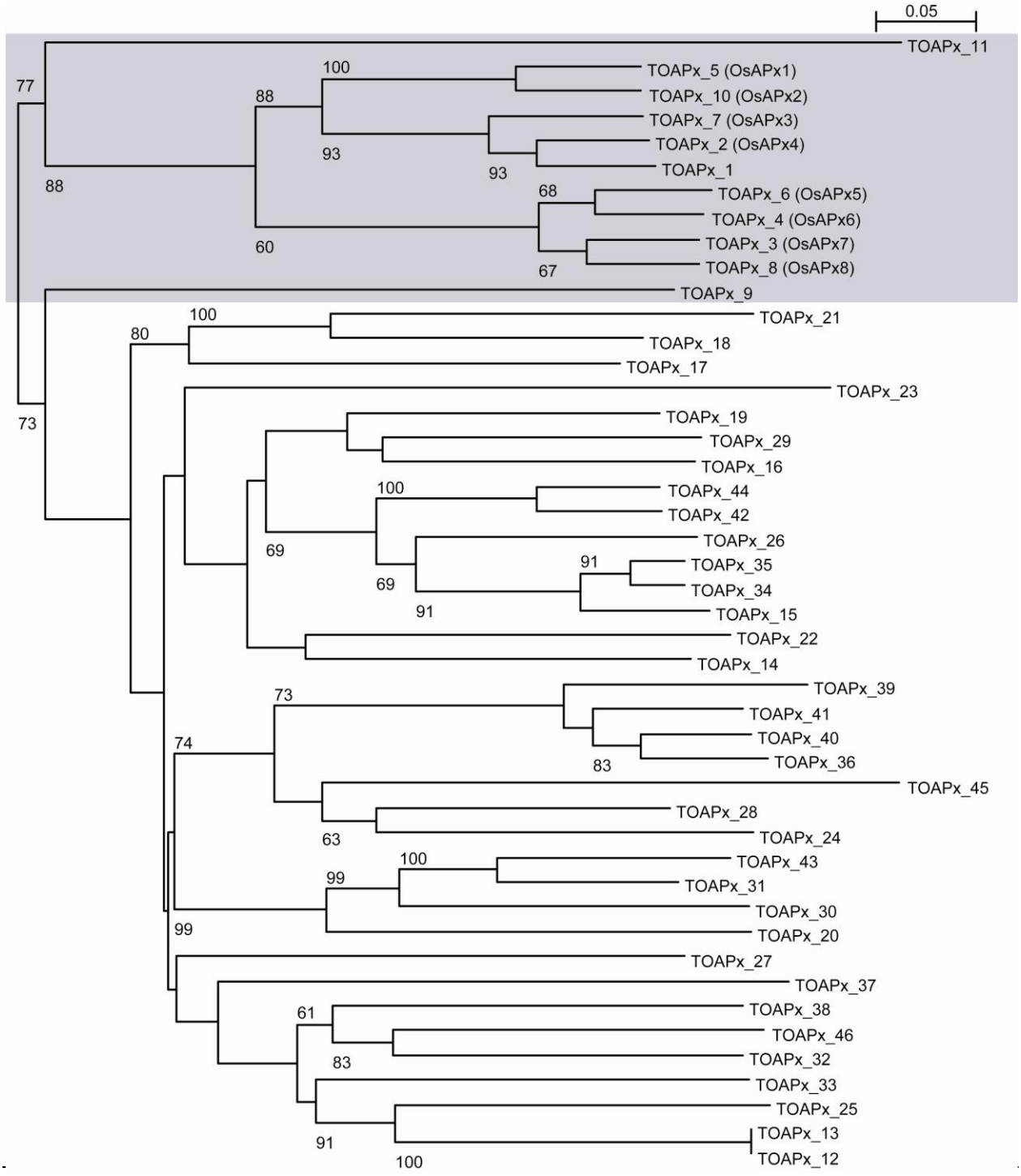


Figure 2.6 An unrooted phylogenetic tree of all rice APx family members predicted by TARGeT. Previously characterized APx gene names are in brackets. The

shaded region contains the true rice APx homologs. Bootstrap values greater than 70 are shown.

2 Searching for APx gene family members in maize and sorghum. To further evaluate the cross species search ability of TARGeT, I searched for APx gene families in maize and sorghum, using the same query that was used to search for rice APx genes. The reasons for choosing maize and sorghum are as follows. First, at the time of the final analysis for this study, the available maize and sorghum sequences were incomplete. Maize is being sequenced using a BAC by BAC approach while sorghum was sequenced using a whole genome shotgun approach. As such, they are more representative of the available genomic databases than the complete rice sequence. Second, search results of maize and sorghum can be compared with the rice and *Arabidopsis* output. Finally, the APx gene families in maize and sorghum have not as yet been characterized.

I identified 11 APx homologs in maize and 9 in sorghum (**Supplemental Figure 2 to 5**). To get a comprehensive view of the APx family in plants, I produced a phylogenetic tree with MEGA (Tamura *et al.* 2007), using the published APx data from *Arabidopsis* and the data predicted by TARGeT for rice, maize and sorghum (**Figure 2.7**). APx gene homologs are clustered into 5 main clades (labelled A-E) with members from all species, indicating several ancient duplications preceding species divergence. The putative new rice APx homologs TOAPx_1, TOAPx_9 and TOAPx_11 are in clades

B, D and E, respectively. Except for two maize homologs in clade D, there is only one representative for each species in clades D and E. This may be due to the effect of gene dosage balance on these two clades (Liang *et al.* 2008; Papp *et al.* 2003; Qian and Zhang 2008). In addition to the main clades, there are several putative orphan clades that are missing genes from one or more species. This may be due to either gene lose or insufficient sequence data.

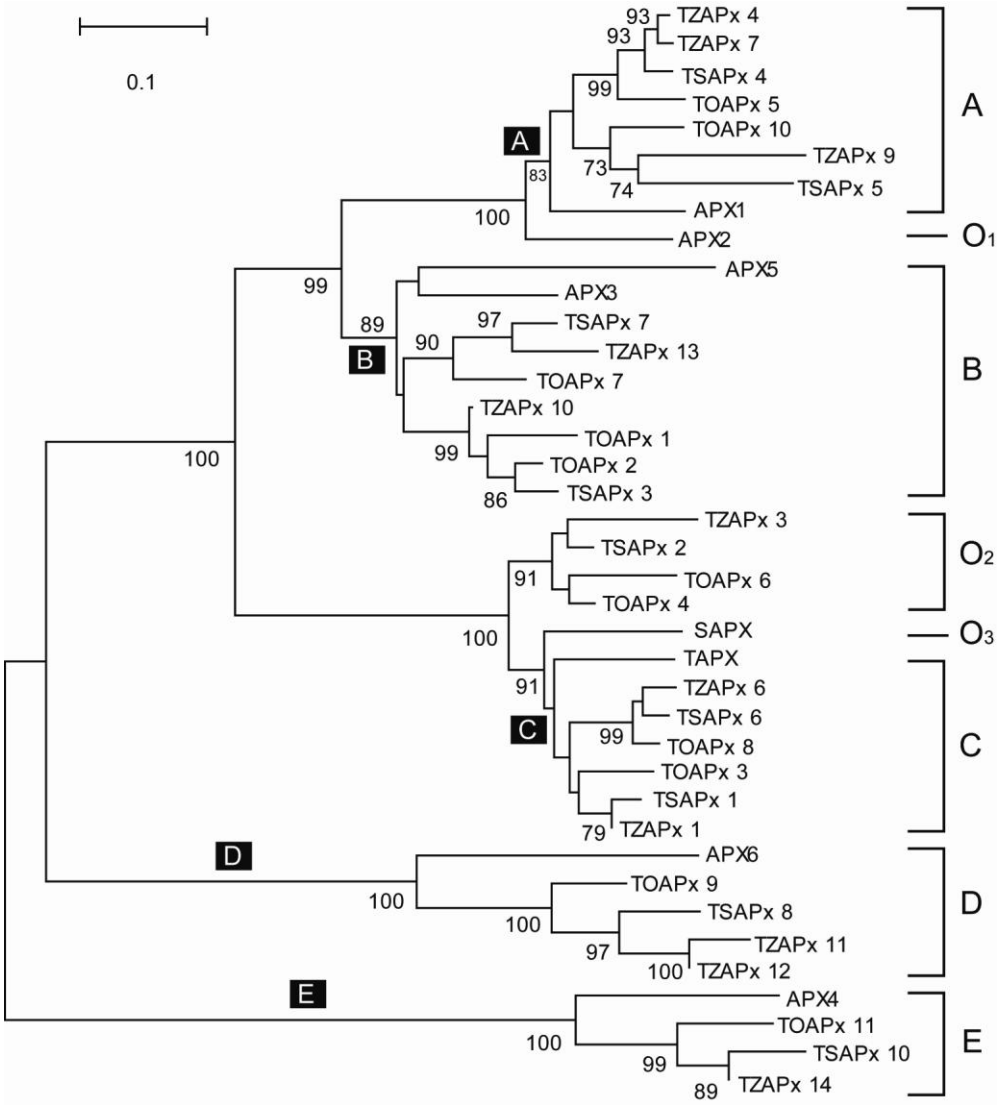


Figure 2.7 An unrooted phylogenetic tree of the APx homologs of rice, maize, sorghum and *Arabidopsis*. This tree was generated with MEGA version4 using the neighbor jointing method with pairwise deletion and p-distance. Five main clades are labeled from A to E. A main clade is defined as a minimal group of homologs that can be found in all species. The remaining homologs are classified into orphan clades O1-O3. Bootstrap values higher than 70 are shown.

3 Searching DNA TE Families in rice. TARGeT is a powerful tool for rapid TE identification, characterization and phylogenetic analysis. I have illustrated this by using TARGeT to search for TEs in the rice genome using as query conserved transposase sequences from five DNA TE superfamilies. The queries were constructed from known TE protein sequences that were downloaded from Repbase (Jurka *et al.* 2005) and additional sequences annotated as part of another study (data not shown). Here I focus on the TARGeT results for the *Tc1/Mariner* superfamily because it has been well annotated and characterized in rice.

The Tc1*/mariner* superfamily is widespread in plant and animal genomes (Feschotte and Wessler 2002). A previous study (Feschotte and Wessler 2002) annotated 34 coding *Mariner*-like Elements (MLEs) from two partially sequenced rice genomes (14 from the *indica* database and 20 from the *japonica* database). Here, I used TARGeT to search the complete *japonica* database and, in about one minute, generated a phylogenetic tree that was consistent with that of Feschotte *et al* (2002). TARGeT successfully retrieved the 20 MLEs reported in the previous study and, in addition, detected 27 new MLEs (**Figure 2.8**).

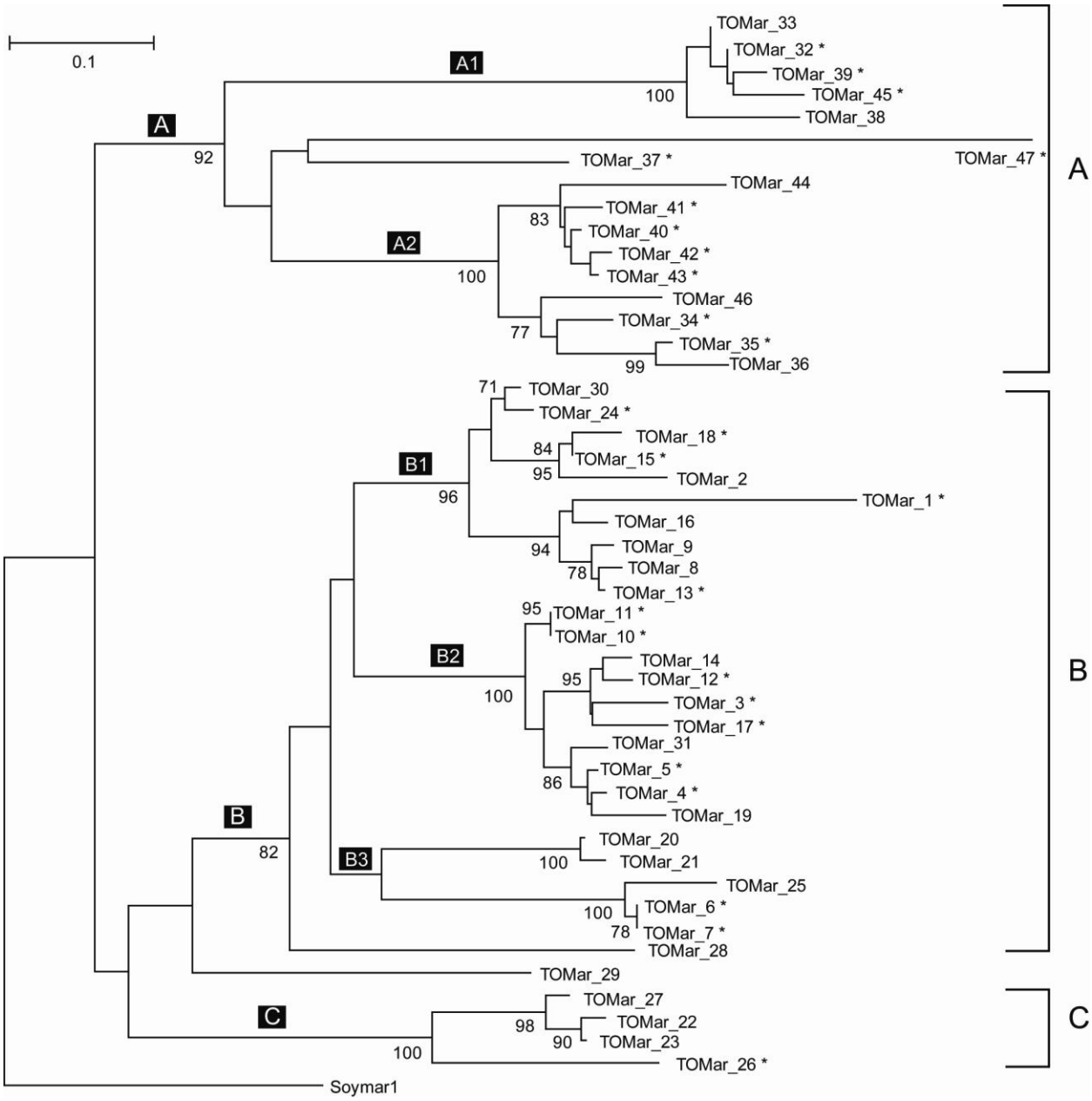


Figure 2.8 A rooted phylogenetic tree of predicted rice Tc1/*mariner* transposases. Three clades, (A, B and C) are defined using the phylogenetic tree generated by Feschotte *et al* (2002). Elements noted by an asterisk are new transposases predicted by TARGeT. *Soymar1* was used as an outgroup and the tree was rooted manually using TreeView. Bootstrap values greater than 70 are shown.

4 Evaluating of the speed of TARGeT. Many factors can affect the speed of TARGeT, such as the number and length of the query sequences, the gene/TE family size, the database size and the number of exons. Other issues that affect the run time include the server hardware and current usage. In addition, because TARGeT is entirely web based, upload and download times vary from user to user. For the gene or TE families that were analyzed in this study, I calculated the average time for each search as an average of 10 independent runs. For example, TARGeT took about 1.2, 2.5 and 6.8 minutes to complete the searches of the APx gene family in rice, sorghum and maize, respectively. The search of the rice Tc1/*mariner* superfamily took about 1 minute to complete.

5 Comparison of TARGeT with similar programs. Two other pipelines, GFScan (Xuan *et al.* 2002) and FGF (Zheng *et al.* 2007), can also retrieve and characterize gene families from genomic databases. GFScan searches for gene family members with the representative genomic DNA motif while FGF performs TBLASTN search followed by GeneWise and phylogenetic analysis. Here I briefly compare the features and performance of TARGeT with these two pipelines.

5.1 TARGeT versus GFScan. The cross-species searching ability of GFScan was previously tested by using a human query sequence to retrieve carbonic anhydrases (CA) genes from the mouse genome (Xuan *et al.* 2002). GFScan was able to identify only 5 of the 11 known CA genes along with 2 putative new CA genes in the available mouse genome sequence. The authors stated that this discrepancy was due to the large difference between the human and mouse motifs. I did a similar search

using TARGeT for CA genes in the mouse. Because there is no record of the version of the mouse genomic database used in the GFScan paper, I chose the latest version of the reference data (October 18, 2006) from Genbank. A query composed of 14 protein sequences from fourteen known human CA genes was constructed. TARGeT found 17 out of 18 known CA genes (data in 2008) in mouse using default parameters, and the 18th was identified after the MMP cut-off was reduced from 0.7 to 0.5.

5.2 TARGeT versus FGF. Direct comparison between the results of TARGeT and FGF proved difficult. First, the FGF server is often not available. Second, TARGeT and FGF use different local databases. I ran TARGeT with the queries that were used in the paper describing FGF. Using a peptidylprolyl isomerase Cyp2 gene (AK061894, GI: 115443875) as query to search against the rice database with default parameters, TARGeT found 6 more putative homologs than FGF (**Supplemental Figure 6**). I also found one possible mistake in the result of FGF: it identified two overlapping homologs, AK061894_chr06 and AK061894_chr06, while there is no such overlap in the result of TARGeT. Using Hsp90 (GI: 40254816) as the query to search against the human database, both FGF and TARGeT found 15 homologs (**Supplemental Figure 7**).

Discussion

To date, most gene family search programs can only retrieve homologs from protein sequence databases. More commonly, BLAST has been widely used to search genomic sequence databases. However, manual retrieval of homolog sequences from BLAST outputs requires a great deal of time. This is especially true for large gene or TE families. TARGeT is particularly useful if one wants to quickly retrieve and characterize gene families from DNA databases, especially when a newly sequenced genome is available. TARGeT uses a Perl program named PHI that automatically retrieves homolog sequences from BLAST outputs. In addition, TARGeT can do multiple alignment and phylogenetic analysis with the retrieved homolog sequences. Speed is another major advantage of TARGeT. As demonstrated in this report, TARGeT can routinely retrieve and characterized gene family homologs, including TEs, from plant and animal genome sequences on the order of minutes.

Although TARGeT shares similarity with homology-based TE annotation tools like RepeatMasker (Unpublished, http://repeatmasker.org), there are some important differences. First, instead of showing each fragmented match as RepeatMasker does, TARGeT tries to identify homologs that are long enough for phylogenetic tree estimation. A fragmented TE can be identified as long as the sum length of its fragments satisfies the MMP to the query. As such, using the same query and databases, the number of homologs identified by TARGeT is usually lower than the hit number found by RepeatMasker. Second, when there are no repeat libraries available for a particular species, RepeatMasker gives the user the option of performing a

BLASTX search to annotate coding regions of TEs in the submitted sequences. In contrast, TARGeT uses a TBLASTN search to identify coding regions from the whole genomic database. Finally, RepeatMasker lacks most of the functionality that is provided by TARGET including the generation of phylogenetic tree and gene structure figures.

When used to search genomic databases, protein sequences queries can efficiently detect distantly related homologs even when their DNA sequences cannot be aligned. Based on our experience, TBLASTN can detect sequences with identities as low as 25% to the query (data not shown). Comparison of the results of TARGeT, FGF and GFScan show that TARGeT retrieved more homologs. To further improve TARGeT's ability to identify distantly related homologs, I are planning to optimize matrix and BLAST parameters (such as gap penalties).

Using multiple queries can also increase the chances of finding additional gene family homologs. TARGeT can accept multiple queries at one time. Although more than one query may hit one homolog, a unique feature of TARGeT is that it can select the one that has the best match to the homolog.

When there is too much sequence divergence between a homolog sequence and the query, the homolog may not be found by TARGeT. However, TARGeT may still provide a clue for users to find them. For most homologs where HSPs are inadequate to meet the MMP cut-off value, they may still have short matches to the query in highly conserved regions. In this case, the file containing the BLAST HSPs that do not meet the qualified homolog cut-off would be valuable. Inspecting this file may give users a

reason why TARGeT failed to detect some homologs and help users design new queries to find additional homologs.

TARGeT uses two approaches to separate closely related gene families. Because there is no absolute similarity cut-off among genes that are within or between families, closely related gene families may be retrieved, under certain circumstances, with the target gene family. This is often the case when the query is short, such as a domain sequence. An efficient way to separate closely related gene families is using phylogenetic analysis because homologs from the same family tend to cluster on a phylogenetic tree into the same clade (**Figure 6 - 8**). However it may not be obvious which clade represents the homologs of interest. In other situations the phylogenetic relationships between the homologs may be ambiguous when the root is unknown.

To overcome these limitations, TARGeT displays the gene structure of each homolog and their sequence similarity to the queries. Because different gene families often have distinct gene structures, homologs that have high sequence similarity to the queries and also have similar gene structures can be easily identified as members of the target gene family. For example, the homologs in the shaded clade in **Figure 6** have higher sequence similarity to the query sequence than the homologs in the other clade, indicating that they are APx homologs. A determination of whether TOAPx_9 and TOAPx_11 belong to the shaded clade requires the gene structure comparison provided by TARGeT (**Supplemental Figure 1**) because the (unrooted) phylogenetic tree alone does not provide sufficient information.

Acknowledgements

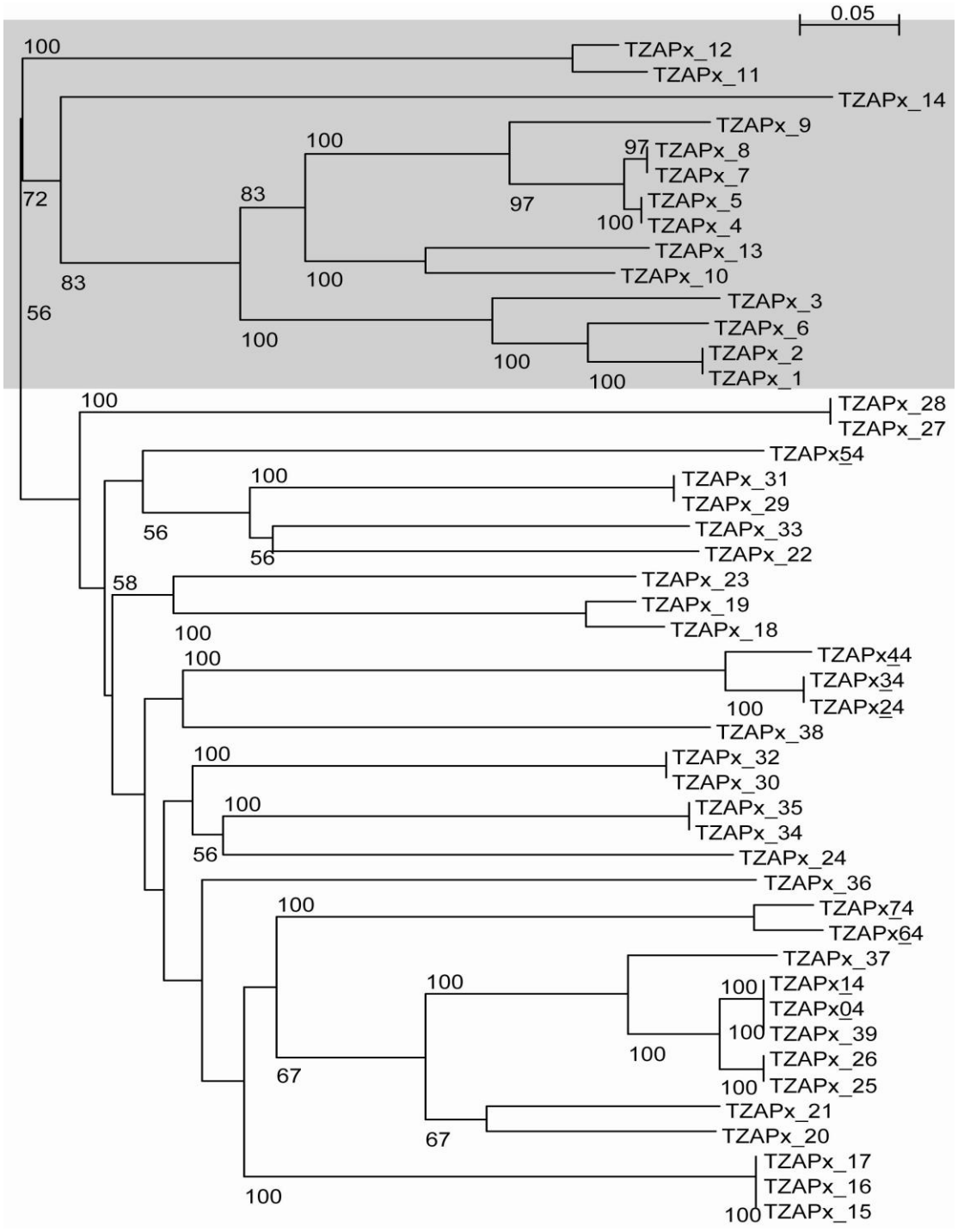
I thank Drs. Hongyan Shan, Jim Leebens-Mack, Russell Malmberg and Yaowu Yuan for reading the manuscript and for valuable discussions. This work was supported by National Science Foundation (Grant DBI-0607123) and the Howard Hughes Medical Institute (Grant 52005731) to SRW.

Supplementary figures

Supplementary Figure 1 Rice APx homologs predicted and annotated by TARGeT using *Arabidopsis* APx genes as query. See the legend of **Figure 5** for more details.

Supplementary Figure 2 Maize APx homologs predicted and annotated by TARGeT using *Arabidopsis* APx genes as query. See the legend of Figure 5 for more details.

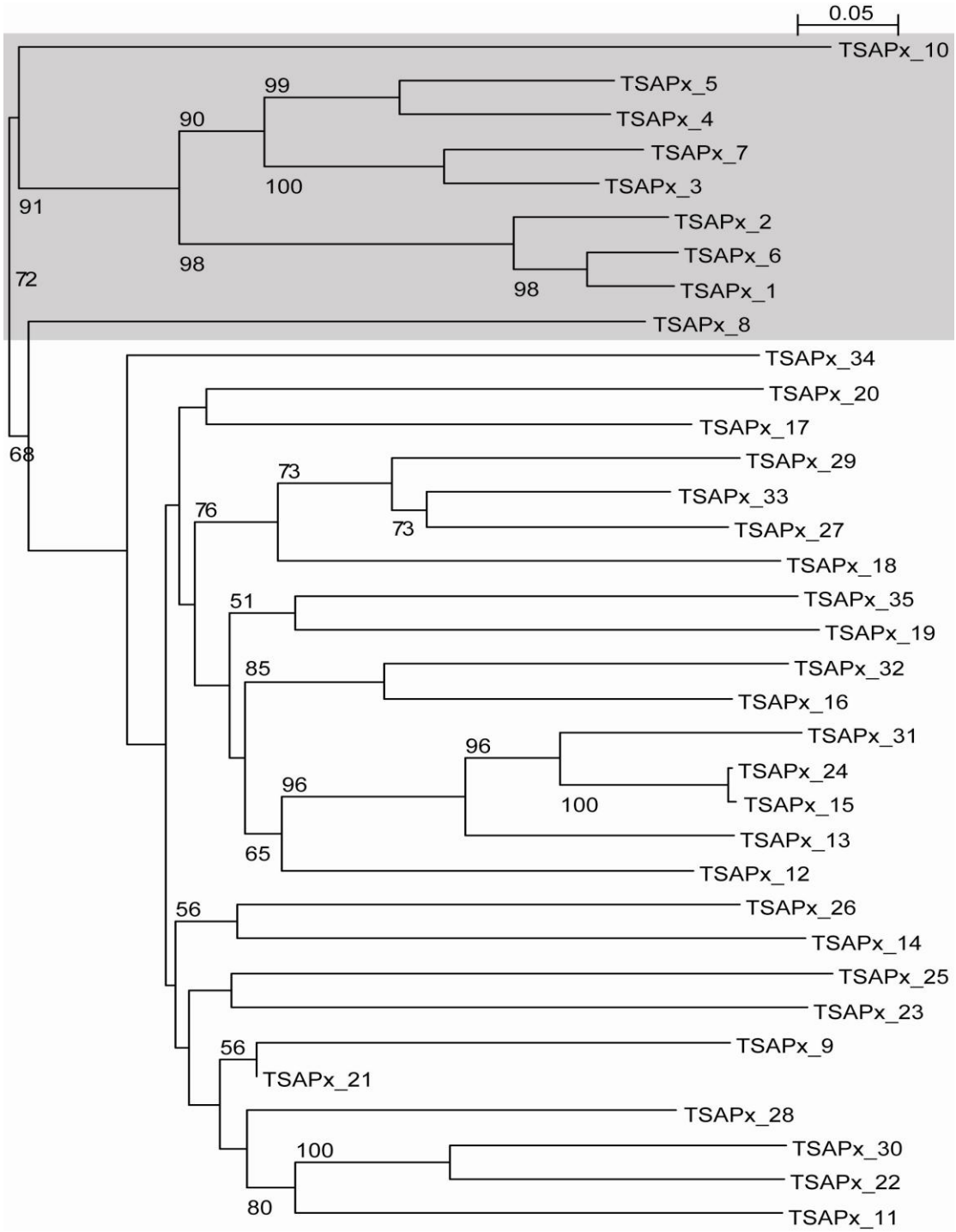
Note that overlaps between BACs may result in artifactual homologs. For instance, TZAPx_1 and TZAPx2 are, in fact, the same gene localized on two overlapping BACs.



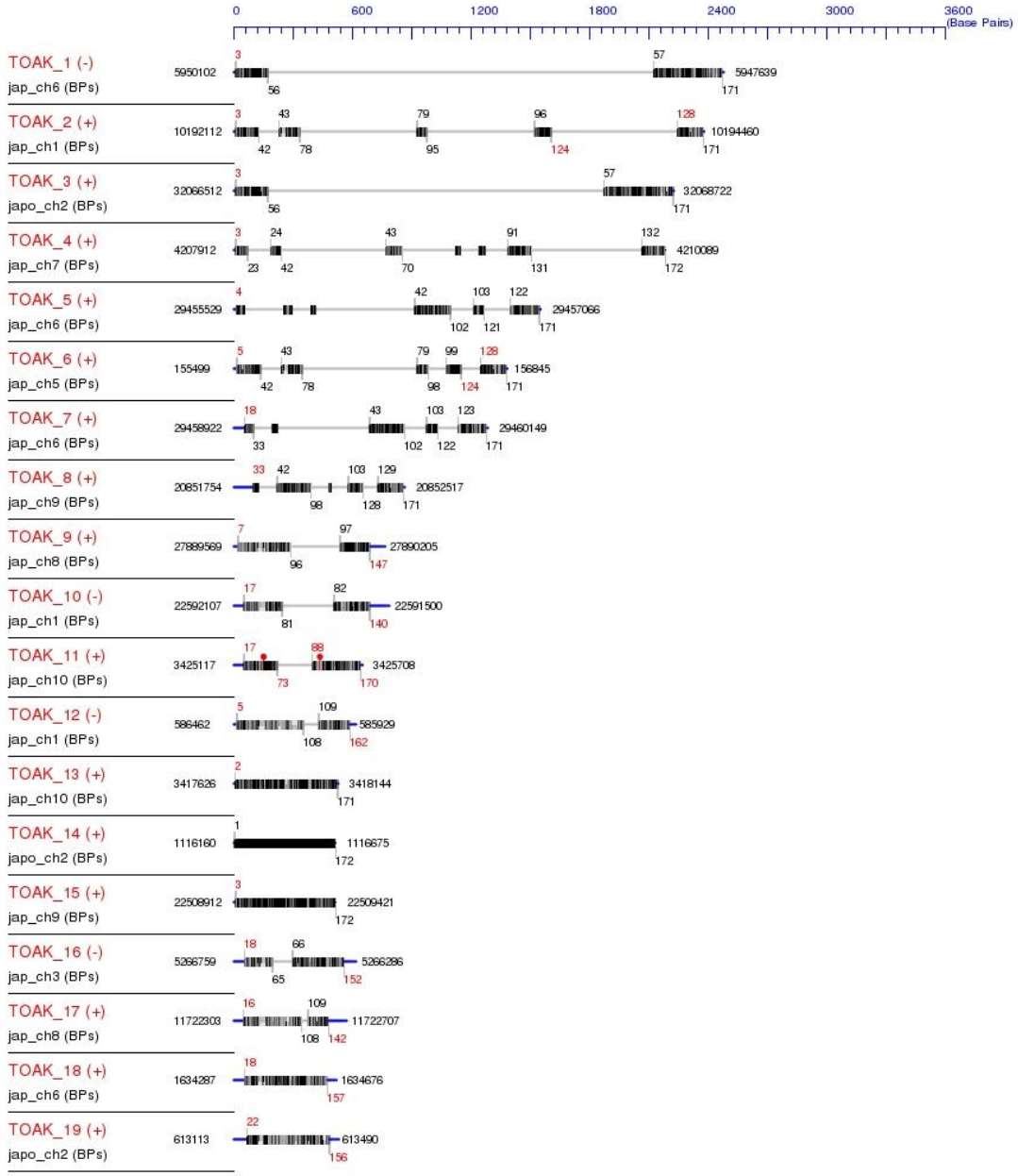
Supplementary Figure 3 Phylogenetic tree of maize APx homologs predicted by TARGeT. See the legend of Figure 6 for more details.


0
900
1800
2700
3600
4500
5400
(Base Pairs)
TSAPx_1 (-)
chromosome_6 (BPs)
46255511
46251728
TSAPx_2 (+)
chromosome_8 (BPs)
6062986
6066305
TSAPx_3 (+)
chromosome_7 (BPs)
59856512
59859775
TSAPx_4 (-)
chromosome_1 (BPs)
62243295
62240603
TSAPx_5 (-)
chromosome_2 (BPs)
77865634
77863005
TSAPx_6 (-)
chromosome_4 (BPs)
52120082
52117514
TSAPx_7 (+)
chromosome_6 (BPs)
3609102
3611620
TSAPx_8 (-)
chromosome_7 (BPs)
62400141
62398150
TSAPx_9 (-)
chromosome_5 (BPs)
1021130
1019371
TSAPx_10 (+)
chromosome_6 (BPs)
56405358
56407016
TSAPx_11 (+)
chromosome_3 (BPs)
73841071
73841955
TSAPx_12 (-)
chromosome_9 (BPs)
5614583
5613762
TSAPx_13 (+)
chromosome_4 (BPs)
61119999
61120793
TSAPx_14 (+)
chromosome_9 (BPs)
45553045
45553821
TSAPx_15 (-)
chromosome_9 (BPs)
3048828
3048067
TSAPx_16 (-)
chromosome_2 (BPs)
74612486
74611734
TSAPx_17 (+)
chromosome_7 (BPs)
1266075
1266821
TSAPx_18 (+)
chromosome_2 (BPs)
461120
461863
TSAPx_19 (+)
chromosome_1 (BPs)
3356609
3357352
TSAPx_20 (-)
chromosome_10 (BPs)
28471037
28470312
TSAPx_21 (-)
chromosome_10 (BPs)
43072107
43071388
TSAPx_22 (-)
chromosome_6 (BPs)
61788574
61787861
TSAPx_23 (+)
chromosome_1 (BPs)
49981442
49982146
TSAPx_24 (-)
chromosome_5 (BPs)
5071233
5070532
TSAPx_25 (-)
chromosome_1 (BPs)
69935763
69935077
TSAPx_26 (-)
chromosome_6 (BPs)
48860068
48859382
TSAPx_27 (+)
chromosome_10 (BPs)
16856275
16856961
TSAPx_28 (+)
chromosome_3 (BPs)
11090913
11091518
TSAPx_29 (+)
chromosome_2 (BPs)
446514
447164
TSAPx_30 (-)
chromosome_6 (BPs)
61806997
61806347
TSAPx_31 (+)
chromosome_9 (BPs)
3064855
3065499
TSAPx_32 (-)
chromosome_2 (BPs)
74620160
74619519
TSAPx_33 (+)
chromosome_2 (BPs)
426829
427452
TSAPx_34 (+)
chromosome_4 (BPs)
67967362
67967964
TSAPx_35 (-)
chromosome_1 (BPs)
57864563
57863979

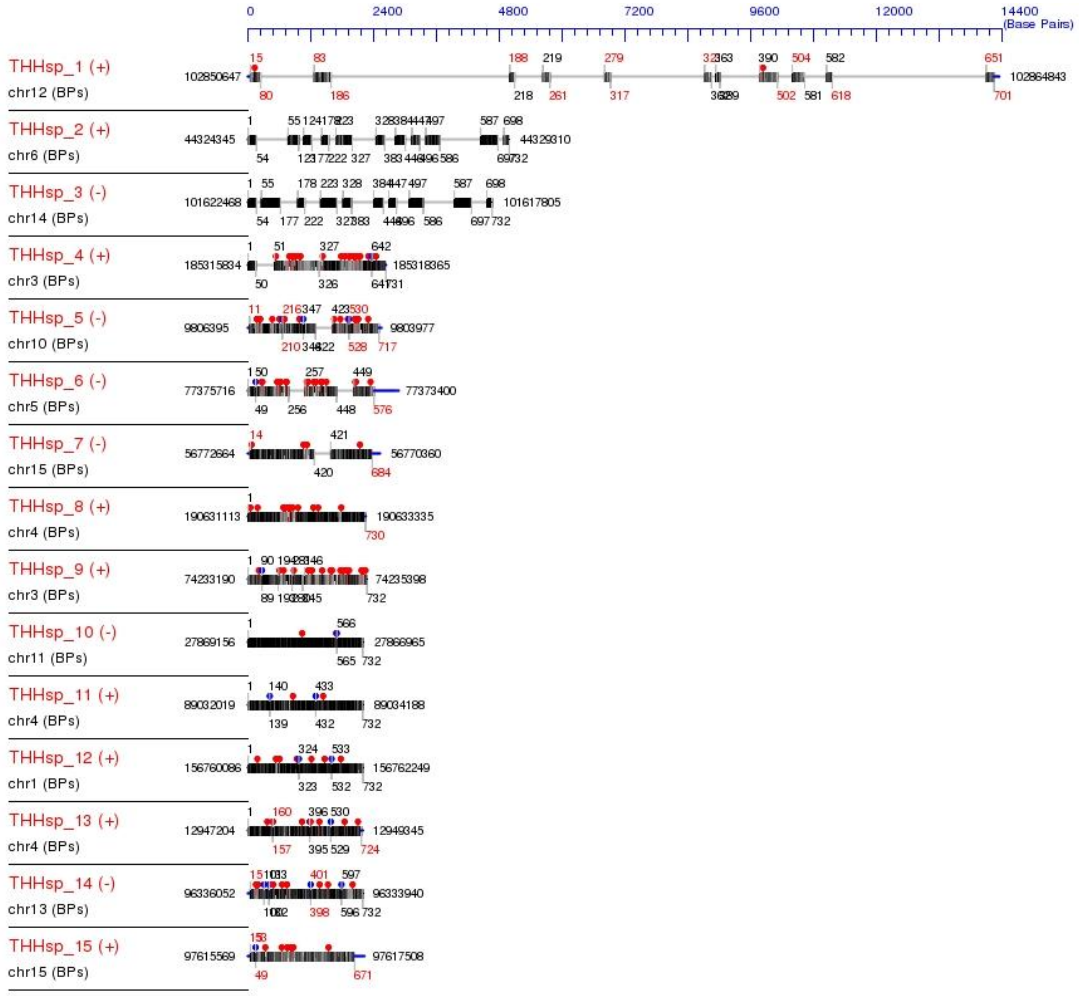
Supplementary Figure 4 Sorghum APx homologs predicted and annotated by TARGeT, using *Arabidopsis* APx genes as query. See the legend of Figure 5 for more details.


0.05
TSAPx_10
99
TSAPx_5
TSAPx_4
90
TSAPx_7
100
TSAPx_3
91
TSAPx_2
98
TSAPx_6
72
98
TSAPx_1
TSAPx_8
TSAPx_34
TSAPx_20
TSAPx_17
68
TSAPx_29
73
TSAPx_33
76
73
TSAPx_27
TSAPx_18
51
TSAPx_35
TSAPx_19
85
TSAPx_32
TSAPx_16
TSAPx_31
96
TSAPx_24
96
100
TSAPx_15
TSAPx_13
65
TSAPx_12
56
TSAPx_26
TSAPx_14
TSAPx_25
TSAPx_23
56
TSAPx_9
TSAPx_21
TSAPx_28
100
TSAPx_30
TSAPx_22
80
TSAPx_11

Supplementary Figure 5 Phylogenetic tree of sorghum APx homologs predicted by TARGeT. Bootstrap values greater than 50 are shown. See the legend of Figure 6 for more details.



Supplementary Figure 6 Homologs of AK061894 in rice as predicted by TARGeT. Premature stop codons and frameshifts are marked by red and blue circles, respectively. Query is AK061894. See the legend of Figure 5 for more details.



Supplementary Figure 7 Hsp90 homologs in the human genome as predicted by TARGeT. The query was Hsp90. See the legend of Figure 5 for more details. Premature stop codons and frameshifts are marked by red and blue circles, respectively.

Supplementary Figure 8 Mariner-like elements in rice, detected by TARGeT, using known Mariner transposase sequences as query. Premature stop codons and

frameshifts are marked by red and blue circles, respectively. See the legend of Figure 5 for more details.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Arigon, A.M., Perriere, G., and Gouy, M. 2006. HoSeqI: automated homologous sequence identification in gene family databases. *Bioinformatics* **22:** 1786-1787.

Bailey, J.A., Liu, G., and Eichler, E.E. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* **73:** 823-834.

Birney, E., Clamp, M., and Durbin, R. 2004. GeneWise and Genomewise. *Genome Res* **14:** 988-995.

Brinkman, F.S., Wan, I., Hancock, R.E., Rose, A.M., and Jones, S.J. 2001. PhyloBLAST: facilitating phylogenetic analysis of BLAST results. *Bioinformatics* **17:** 385-387.

Burks, C., Fickett, J.W., Goad, W.B., Kanehisa, M., Lewitter, F.I., Rindone, W.P., Swindell, C.D., Tung, C.S., and Bilofsky, H.S. 1985. The GenBank nucleic acid sequence database. *Comput Appl Biosci* **1:** 225-233.

Cheung, J., Wilson, M.D., Zhang, J., Khaja, R., MacDonald, J.R., Heng, H.H., Koop, B.F., and Scherer, S.W. 2003. Recent segmental and gene duplications in the mouse genome. *Genome Biol* **4:** R47.

Dayhoff, M.O. 1976. The origin and evolution of protein superfamilies. *Fed Proc* **35:** 2132-2138.

Dehal, P. and Boore, J.L. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* **3:** e314.

Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* **14:** 755-763.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32:** 1792-1797.

Eichler, E.E. 2001. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet* **17:** 661-669.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30:** 1575-1584.

Feschotte, C. and Wessler, S.R. 2002. Mariner-like transposases are widespread and diverse in flowering plants. *Proc Natl Acad Sci U S A* **99:** 280-285.

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L. et al. 2008. The Pfam protein families database. *Nucleic Acids Res* **36:** D281-288.

Frank, R.L., Mane, A., and Ercal, F. 2006. An Automated Method for Rapid Identification of Putative Gene Family Members in Plants. *BMC Bioinformatics* **7 Suppl 2:** S19.

Frickey, T. and Lupas, A.N. 2004. PhyloGenie: automated phylome generation and analysis. *Nucleic Acids Res* **32:** 5231-5238.

Gertz, E.M., Yu, Y.K., Agarwala, R., Schaffer, A.A., and Altschul, S.F. 2006. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol* **4:** 41.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H. et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* **296:** 92-100.

Hanekamp, K., Bohnebeck, U., Beszteri, B., and Valentin, K. 2007. PhyloGena--a user-friendly system for automated phylogenetic annotation of unknown sequences. *Bioinformatics* **23:** 793-801.

Heger, A. and Holm, L. 2000. Towards a covering set of protein family profiles. *Prog Biophys Mol Biol* **73:** 321-337.

Hurles, M. 2004. Gene duplication: the genomic trade in spare parts. *PLoS Biol* **2:** E206.

Initiative, A.G. 2000a. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408:** 796-815.

Initiative, T.A.G. 2000b. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408:** 796-815.

Jiang, N., Bao, Z., Zhang, X., Eddy, S.R., and Wessler, S.R. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431:** 569-573.

Jordan, G.E. and Piel, W.H. 2008. PhyloWidget: web-based visualizations for the tree of life. *Bioinformatics* **24:** 1641-1642.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110:** 462-467.

Karsch-Mizrachi, I. and Ouellette, B.F. 2001. The GenBank sequence database. *Methods Biochem Anal* **43:** 45-63.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664.

Koski, L.B. and Golding, G.B. 2001. The closest BLAST hit is often not the nearest neighbor. *J Mol Evol* **52:** 540-542.

Koszul, R., Caburet, S., Dujon, B., and Fischer, G. 2004. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J* **23:** 234-243.

Lander, E.S. Linton, L.M. Birren, B. Nusbaum, C. Zody, M.C. Baldwin, J. Devon, K. Dewar, K. Doyle, M. FitzHugh, W. et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Li, H., Coghlan, A., Ruan, J., Coin, L.J., Heriche, J.K., Osmotherly, L., Li, R., Liu, T., Zhang, Z., Bolund, L. et al. 2006. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res* **34:** D572-580.

Li, W.H., Gu, Z., Wang, H., and Nekrutenko, A. 2001. Evolutionary analyses of the human genome. *Nature* **409:** 847-849.

Liang, H., Plazonic, K.R., Chen, J., Li, W.H., and Fernandez, A. 2008. Protein under-wrapping causes dosage sensitivity and decreases gene duplicability. *PLoS Genet* **4:** e11.

Meyers, B.C., Tingey, S.V., and Morgante, M. 2001. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res* **11:** 1660-1676.

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M.J. et al. 2005. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res* **33:** D284-288.

Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A., and Rafalski, A. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet* **37:** 997-1002.

Page, R.D. 1996. TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* **12:** 357-358.

Papp, B., Pal, C., and Hurst, L.D. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424:** 194-197.

Passardi, F., Theiler, G., Zamocky, M., Cosio, C., Rouhier, N., Teixera, F., Margis-Pinheiro, M., Ioannidis, V., Penel, C., Falquet, L. et al. 2007. PeroxiBase: the peroxidase database. *Phytochemistry* **68:** 1605-1611.

Pearson, W.R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol* **183:** 63-98.

Qian, W. and Zhang, J. 2008. Gene dosage and gene duplicability. *Genetics* **179:** 2319-2324.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W. et al. 2000. Comparative genomics of the eukaryotes. *Science* **287:** 2204-2215.

Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406-425.

Sicheritz-Ponten, T. and Andersson, S.G. 2001. A phylogenomic approach to microbial evolution. *Nucleic Acids Res* **29:** 545-552.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24:** 1596-1599.

Tchenio, T., Segal-Bendirdjian, E., and Heidmann, T. 1993. Generation of processed pseudogenes in murine cells. *EMBO J* **12:** 1487-1497.

Vanin, E.F. 1985. Processed pseudogenes: characteristics and evolution. *Annu Rev Genet* **19:** 253-272.

Wendel, J.F. 2000. Genome evolution in polyploids. *Plant Mol Biol* **42:** 225-249.

Wolfe, K.H. and Shields, D.C. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387:** 708-713.

Xuan, Z., McCombie, W.R., and Zhang, M.Q. 2002. GFScan: a gene family search tool at genomic DNA level. *Genome Res* **12:** 1142-1149.

Yang, Y.W., Lai, K.N., Tai, P.Y., and Li, W.H. 1999. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. *J Mol Evol* **48:** 597-604.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X. et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* **296:** 79-92.

Zheng, H., Shi, J., Fang, X., Li, Y., Vang, S., Fan, W., Wang, J., Zhang, Z., Wang, W., and Kristiansen, K. 2007. FGF: a web tool for Fishing Gene Family in a whole genome database. *Nucleic Acids Res* **35:** W121-125.

Chapter 3

MITE-Hunter -- a program for discovering miniature inverted-repeat transposable elements (MITEs) from genomic sequences

---

Yujun Han and Susan R. Wessler. 2010.

Submitted to Nucleic Acids Research.

Abstract

Miniature inverted-repeat transposable elements (MITEs) are a special type of class 2 nonautonomous transposable element that are abundant in the noncoding regions of the genes of many plant and animal species. The accurate identification of MITEs has been a challenge for existing programs because they lack coding sequences and, as such, evolve very rapidly. Because of their importance to gene and genome evolution, I developed MITE-Hunter, a program pipeline that can identify MITEs as well as other small class 2 nonautonomous TEs from genomic DNA datasets. The output of MITE-Hunter is composed of consensus TE sequences grouped into families that can be used as a library file for homology-based TE detection programs such as RepeatMasker. MITE-Hunter was evaluated by searching the rice genomic database and comparing the output with known rice TEs. It discovered most of the previously reported rice MITEs (97.6%), and found eleven new elements. MITE-Hunter was also compared with two other MITE discovery programs, FINDMITE and MUST. Unlike MITE-Hunter, neither of these programs can search large genomic datasets including whole genome sequences. More importantly, MITE-Hunter is significantly more accurate than either FINDMITE or MUST as the vast majority of their outputs are false positives.

Introduction

Transposable elements (TEs) reside in most eukaryotic genomes where they contribute to adaptive changes by shaping gene and genome structures (Dooner and Weil 2007; Feschotte *et al.* 2002; Feschotte and Pritham 2007). One of the remarkable discoveries of the past two decades is that the genomes of most higher eukaryotic are composed largely of TEs. For example, sequences derived from TEs make up at least 31% of the genome of dog (*Canis familiaris*), 38% of mouse (*Mus musculus*), 46% of human (*Homo sapiens*) and 85% of maize (*Zea mays* ssp. *mays* L.) (Kirkness *et al.* 2003; Lander *et al.* 2001; Schnable *et al.* 2009; Waterston *et al.* 2002). TEs have structural features and classification systems that serve to distinguish them from simpler repetitive sequences like microsatellite repeats. They are divided into two classes based on the molecule involved in transposition: retrotransposons (class 1) move via a RNA intermediate while DNA is the intermediate of DNA transposons (class 2). In each class, TEs are further divided into superfamilies and families (Wicker *et al.* 2007). In plants, six class 2 superfamilies have been identified thus far: *Tc1/Mariner*, *PIF/Harbinger*, *hAT*, *MULE*, *CACTA* and *Helitron* (Feschotte and Pritham 2007; Wicker *et al.* 2007). With the exception of *Helitrons*, TEs in all superfamilies have terminal inverted repeats (TIRs) and transpose through a cut-and-paste mechanism. Based on whether they can produce functional transposase, TEs are also classified as autonomous or nonautonomous elements.

Miniature inverted-repeat transposable elements (MITEs) are a special type of nonautonomous element that is present in high copy numbers in many eukaryotic genomes. For example, about 56,000 MITEs were identified in sorghum (*Sorghum bicolor*) (Paterson *et al.* 2009), 73,500 in rice (*Oryza sativa*) (Oki *et al.* 2008) and 150,000 in human (Smit and Riggs 1996). Ever since their discovery almost 20 years ago (Bureau and Wessler 1992; Bureau and Wessler 1994), MITEs have been the subject of increasing interest in both plants and animals (Kuang *et al.* 2009; Osborne *et al.* 2006; Tu 2001; Yang *et al.* 2009). Unlike the "traditional" low copy nonautonomous TEs (such as the *Ds* element of maize), MITEs are uniformly short (most < 500 bp) and amplify rapidly from one or a few elements to very high copy numbers (Naito *et al.* 2006). The two largest MITEs families, *Stowaway* and *Tourist*, were found to be members of the *Tc1/Mariner* and the *PIF/Harbinger* superfamilies, respectively (Feschotte *et al.* 2003; Jiang *et al.* 2003; Yang *et al.* 2009; Zhang *et al.* 2001). MITEs have also been reported from the *hAT* and *MULE* superfamilies (Kuang *et al.* 2009; Moreno-Vazquez *et al.* 2005).

While the rapidly expanding database of genomic sequence presents an opportunity to expand the study of MITEs, it also poses a significant challenge to their correct and efficient annotation. Many TE annotation programs have been developed that use one or more of the following computational approaches: 1) homology-based 2) *de novo* 3) polymorphism-based and 4) structure-based (Bergman and Quesneville 2007; Lerat 2009; Surya *et al.* 2008). Homology-based TE annotation is powerful at detecting TEs that share sequence similarity with known elements, but it is inadequate

at identifying full-length or novel TEs. Methods using *de novo* approaches can discover all TEs as long as they have multiple copies. However, the drawback of this approach is that its output is a mixture of TEs from all superfamilies and non-TE repeats. As such, the manual identification and classification of TEs from the output of *de novo* methods is often very tedious and time-consuming. Polymorphism-based approaches can discover new TEs but the output is also a mixture of different types of sequences. More importantly, its application is limited to the comparison of datasets from very closely related species. When compared to the other algorithms, structure-based approaches are very effective at discovering certain TE types like LTR retrotransposons. However, currently available programs are less successful at identifying other TE types like nonautonomous DNA transposons (including MITEs) because they possess few distinguishing structural features.

To date three programs have been developed exclusively to find MITEs: TRANSPO (Santiago *et al.* 2002), FINDMITE (Tu 2001) and MUST (Chen *et al.* 2009). TRANSPO is a homology-based program that requires known MITE sequences. As such it is not effective at finding new MITEs (Lerat 2009). FINDMITE and MUST are structure-based TE discovery programs that can be used to discover new MITEs because they search for common MITE structural features rather than similar sequences. However, because MITEs have only two common structural features, terminal inverted repeats (TIRs) and target site duplications (TSDs), many sequences that are not MITEs are in the outputs of FINDMITE and MUST. Thus, the false positive

rates of these programs are very high and extensive manual curation is required to filter false positives from their output files.

Here I present MITE-Hunter, a program that accurately discovers MITEs as well as other short nonautonomous “cut-and-paste” DNA TEs in genomic datasets including those of whole genomes. To evaluate MITE-Hunter, I compared it with FINDMITE and MUST. I chose the rice genome to evaluate the performance of MITE-Hunter because rice harbors abundant and well-annotated DNA TEs and MITEs (Bureau *et al.* 1996; Jiang *et al.* 2004; Oki *et al.* 2008). In the examples reported in this study, MITE-Hunter missed only two known rice MITEs and discovered eleven previously unknown elements. Compared to FINDMITE and MUST, MITE-Hunter has a much lower false positive rate and the output is easier to be checked and classified. MITE-Hunter and related programs can be freely downloaded at http://target.iplantcollaborative.org/.

Materials and Methods

The MITE-Hunter Pipeline. MITE-Hunter is a UNIX program pipeline composed mainly of Perl scripts. Given genomic sequences as the input data, MITE-Hunter identifies class 2 nonautonomous TEs and produces outputs of consensus sequences classified into families. MITE-Hunter can use multiple processers (default 5 CPUs). The MITE-Hunter pipeline has five main steps that are summarized in **Figure 3.1**: (I) identify TE candidates through a structure-based approach, (II) identify and filter false positives using an approach based on the pairwise sequence alignment (PSA), (III) generate exemplars, (IV) identify and filter false positives using an approach based on the multiple sequence alignment (MSA), generate consensus sequences and predict TSDs and (V) group consensus sequences into families. Details of each step are presented in the results section.

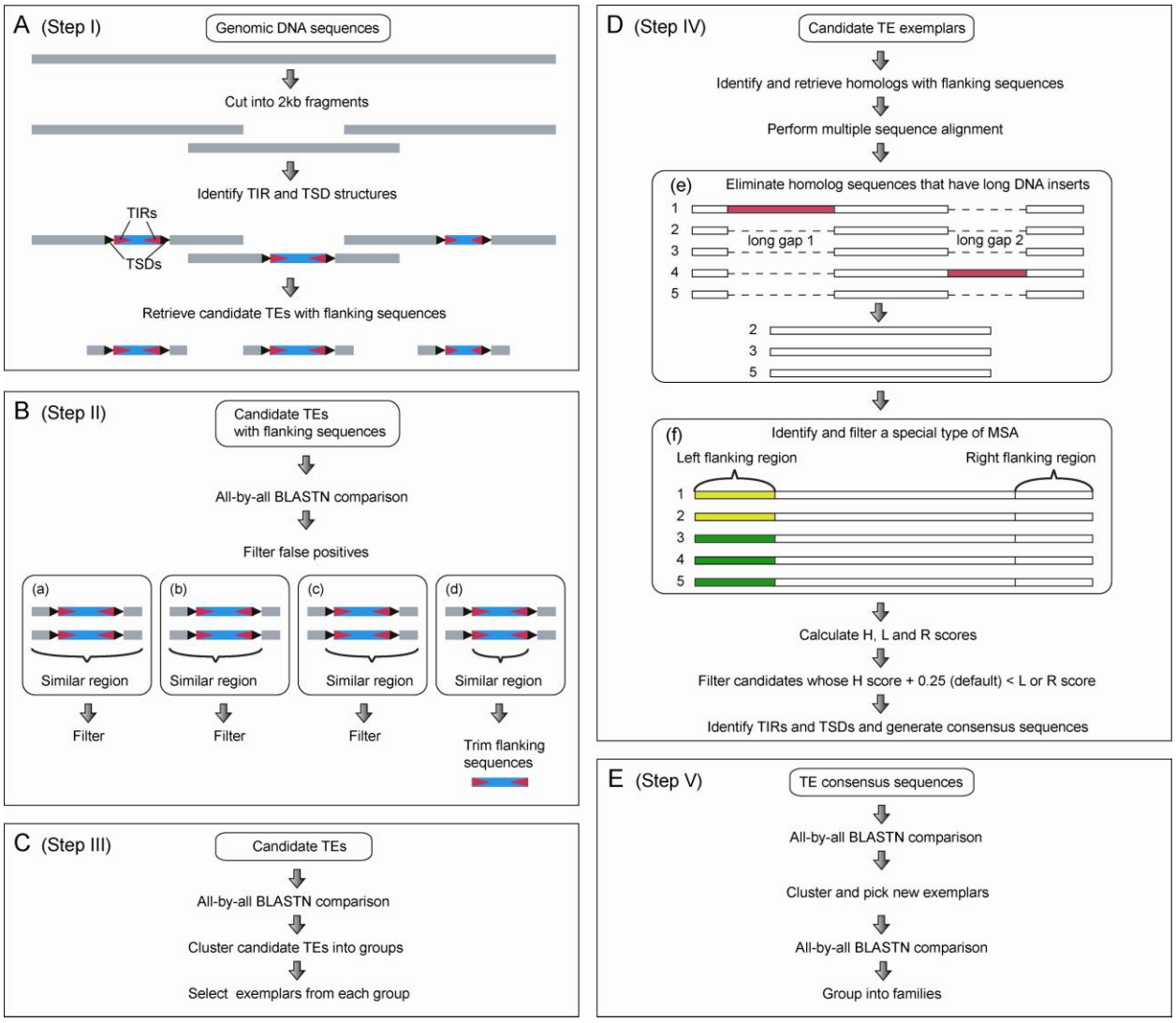


Figure 3.1. The five main steps of the MITE-Hunter pipeline. Grey bars are genomic sequences, black and red triangles are TSDs and TIRs, respectively, blue bars are predicted TEs, white bars are homolog sequences, dashed lines are gaps and yellow bars are sequences that are similar to each other but not to those represented by green bars (and vice versa). A) Identification of candidate TEs. Three predicted candidate TEs are shown. B) Filtering of false positives based on the PSA. Four types of alignments are shown (a - d). Except for the candidates in (d), all the others are

filtered as false positives. C) Selection of TE exemplars. D) Filtering of false positives based on the MSA, predicting TSDs and generating consensus sequences. (e) and (f) are two special types of MSA (see text for detail). V) Selecting new exemplars and grouping TEs into families.

Dataset and Programs. The build 5 rice IRGSP/RAP genome sequence was used (Tanaka *et al.* 2008) as was Repbase version 14.02 (Jurka *et al.* 2005) and RepeatMasker 3.26 (AFA *et al.* 2004). TE copy number was calculated using a previously described method (Schnable *et al.* 2009). Pairwise sequences alignment (PSA) used BLAST (Altschul *et al.* 1997) and multiple sequences alignment (MSA) used Muscle (Edgar 2004). All computation was done on a Linux cluster.

Results

MITE discovery in rice. I applied MITE-Hunter to the rice genome with default parameters. MITE-Hunter completed the analysis in about 44 hours. Details of the algorithms and results of each step of MITE-Hunter are presented below.

(I) Identifying all candidates (Figure 3.1A). MITE-Hunter uses genomic sequences as the input data. Long input sequences are first cut into small fragments (default 2 kb) with overlaps (default 500 bp). TE candidates are identified from each fragment sequence as those that have TIR-like structures (default 10 bp with at most 1 bp mismatch) flanked by putative TSDs (2-10 bp; default is TA if TSD length = 2). Because low complexity sequences (LCS) are rare in MITEs but make up many TIR-like and TSD-like structures, TE candidates that have LCS in TIRs or have too many LCS within internal sequences are filtered as follows. First, TIRs that have stretches of tandem 1 to 2 nucleotide units (default >= 8 bp) or have low G+C content (default < 20%) are filtered. Second, a candidate will be filtered if it has too many LCS (default >= 20%) identified by DUST (R. Tatusov and D.J. Lipman, unpublished). Using rice genomic DNA sequences as the input data (~ 380Mb), 629,698 candidate TEs were identified and retrieved together with their flanking sequences (default 60 bp).

(II) Filtering false positives based on the pairwise sequence alignment (PSA) (Figure 3.1B). Candidate TEs and their flanking sequences are submitted to an all-by-all BLASTN comparison (default E value = $1e^{-10}$). To reduce the computational load, candidates are divided into groups based on their length (default interval = 100 bp) and BLASTN is performed separately for each group. From the BLASTN results, single copy candidates are identified and filtered. Of the remaining candidates, only those that share sequence similarity within but not in their flanking regions are retained. Four types of PSAs are shown in **Figure 3.1B** (a - d). In type (a), the similar region extends to both sides. In type (b) and (c), similar regions extend to the left and right of the flanking regions, respectively. Only in type (d) is the similar region within the TIRs and the candidate not filtered as a false positive. Of the 629,698 rice candidates from Step I, 38,617 passed this filter. These candidates were trimmed of their flanking sequences before being sent to the next step.

(III) Identifying TE exemplars (Figure 3.1C). To reduce computational load in the following steps, MITE-Hunter clusters TE candidates based on their sequence similarity and picks one as the exemplar that best characterizes the features of each group. First, the candidates from Step II are subjected to an all-by-all BLASTN comparison. Based on the BLASTN results, candidates are clustered as follows: 1) the candidate that matches most of the others (default matched length percentage >90% and identify >= 80%) is selected as the exemplar, 2) the exemplar and the candidates that it matches are put into one group and will not be sampled again and 3) repeat 1

and 2 until no candidates remain. In this step, of 38,617 TE candidates from Step II, 3,887 exemplars were selected and sent to the next step.

(IV) Filtering false positives using the multiple sequence alignment (MSA), generating consensus sequences and predicting TSDs (Figure 3.1D). Each exemplar identified in Step III is used as a query to perform BLASTN searches of the genomic database. Homologs are identified and retrieved together with their flanking sequences (default 60 bp) by a command line version of TARGeT (Han *et al.* 2009). Candidates that have too few homologs (default >= 3) are filtered because many ultimately prove to be false positives. A MSA is generated using homologs of each exemplar. To reduce the computational load, if there are too many homologs for an exemplar, only the top 35 with the highest BLASTN alignment scores are used. From each MSA, three average identity scores are calculated from the left flanking region (L), homologous region (H) and the right flanking region (R).

$$L = \frac{\sum_{i=1}^{b-1} max_i(S)}{b-1} \qquad H = \frac{\sum_{i=b}^{e} max_i(S)}{e-b+1} \qquad R = \frac{\sum_{i=e+1}^{n} max_i(S)}{n-e}$$

In these equations, *b* and *e* are the beginning and ending positions of homologs in the MSA, *n* is the total length of the MSA, and *S* is the percentage of different nucleotides in each column of the MSA. Candidates whose H score is significantly higher than both L and R scores (default > 0.25) are retained.

Two special situations that can potentially confound the results are addressed in MITE-Hunter. One concerns TE homologs with DNA inserts that cause large gaps in the MSA and significantly lower the H score. In this case, homologs with additional sequences (default > 25 bp) are identified and filtered before calculating the H, L and R

scores. An example is shown in **Figure 3.1(e)**, where the MSA has two long gaps caused by the additional sequences in homologs 1 and 4 (represented by red bars). After filtering homologs 1 and 4, a new MSA is generated using the remaining homolog sequences (2, 3 and 5). The other special situation is that for some MSAs, although the L and R scores are low, a subgroup of flanking sequences is very similar. Based on our experience, most candidates with this type of MSA are false positives. An example is shown in **Figure 3.1(f)**, where the L and R scores are much lower than the H score of the MSA. However, in the left flanking region of this MSA, sequences in 1 and 2 are very similar but are different from the sequences in 3, 4 and 5, which are also similar to each other. To filter this type of false positive, MITE-Hunter calculates the identity between flanking sequences from the MSA. In such cases, the candidate will be filtered if more than 50% of the homologs (default value) share more than 60% identity (default value) in their flanking regions.

For candidates that pass these filters, MITE-Hunter generates consensus sequences and predicts TSDs. In general, consensus sequences better represent homologous TEs than exemplars. While exemplars are selected from real TEs that may have mutations that are different from other homologs, consensus TEs are generated from MSAs and are composed of residues that are most abundant in all of the homologs. MITE-Hunter generates consensus sequences by choosing the most frequent nucleotide from each column (default >= 70%) in the homologous region of the MSA. TSDs are predicted again in this step because it is more accurate to predict TSDs based on MSA than from a single sequence in Step I. Identical sequences flanking each

homolog (default 2-10 bp) are identified from the MSA and the most frequent one (default >= 50%) is recognized as the TSD. Predicted TSDs are useful in the manual classification of MITE-Hunter output into superfamilies. Of the 3,887 TE exemplars from Step III, 2,253 were verified and consensus sequences were generated and sent to the final step.

(V) Identifying new exemplars and grouping into families (Figure 3.1E). To further condense the output, new exemplars are selected from the consensus sequences in Step IV using the same approach as in step III. This step is necessary because after replacing the exemplar sequences with consensus sequences in Step IV, the similarity between many TE consensus sequences satisfies the grouping criteria. Of the 2,253 TE consensus sequences from Step IV, 700 new exemplars were selected and used to execute the all-by-all BLASTN comparison. From the BLASTN results the exemplars were grouped into 446 families based on the 80-80-80 rule (Wicker *et al.* 2007).

Accuracy evaluation of MITE-Hunter. To test the authenticity of the MITE-Hunter output I curated the 700 rice TE consensus sequences **(Figure 3.2)**. Each MSA file was manually analyzed for TIR and TSD structures that are characteristic of class 2 TE superfamilies found in plant genomes. A TE consensus is validated if it has at least three full-length copies and its ends, characterized by TIRs and TSDs, can be recognized from the MSA file. TEs that do not meet these criteria are considered to be false positives. Using these strict parameters, I identified 46 false positives. In addition, eight solo LTRs and four short *Helitrons* were identified and classified as false positives.

These 12 sequences were in the MITE-Hunter output because they coincidentally have TIR-like and TSD-like structures near their ends. After removing these sequences there were 642 consensus TEs remaining from the original 700, resulting in a false positive rate of 8.3% [(46 + 8 + 4) / 700].

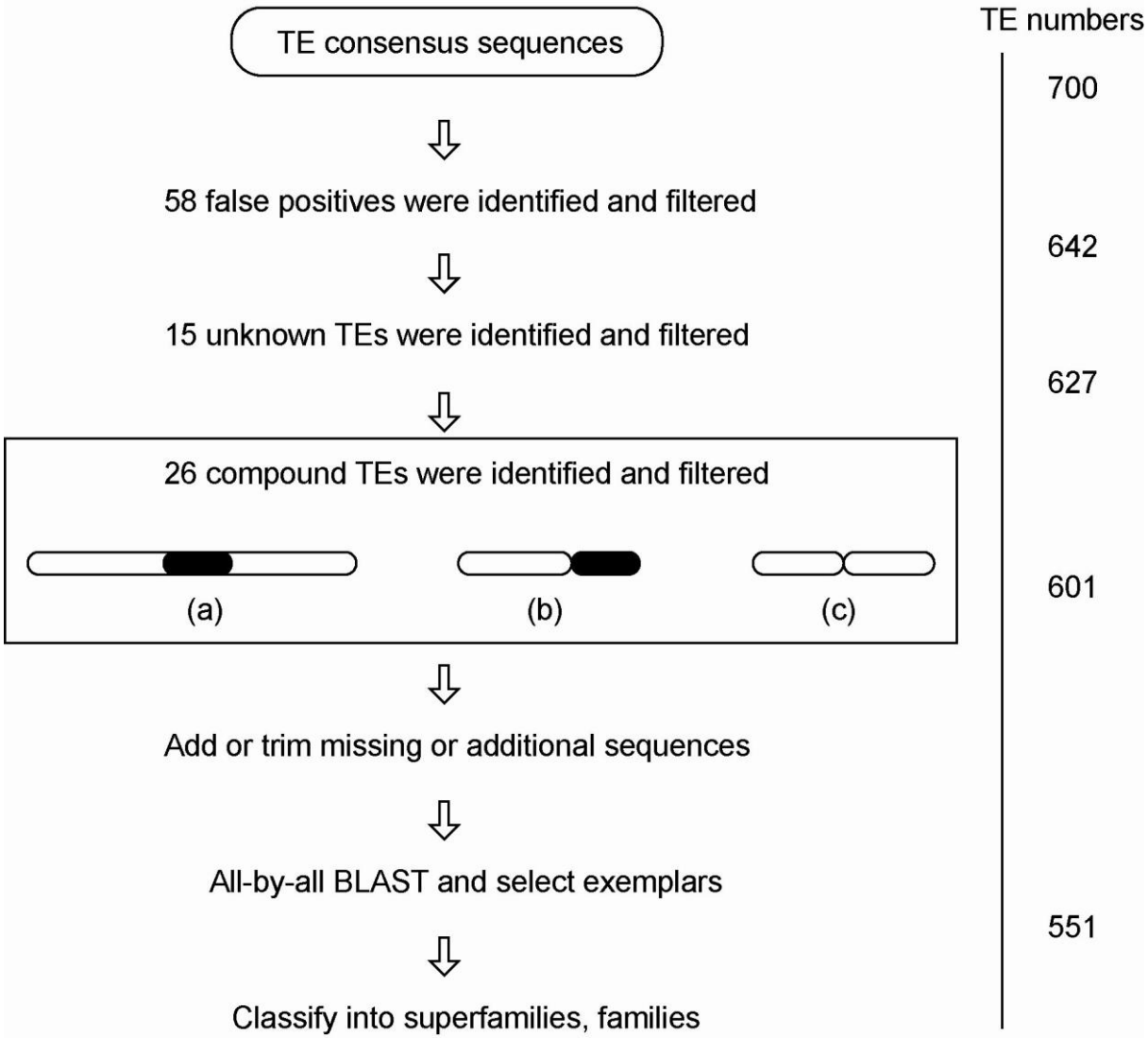


Figure 3.2. Flowchart of the manual curation of rice Class 2 nonautonomous TEs from MITE-Hunter output. The authentication process began with 700 consensus TEs and was reduced by the number shown for each step. The numbers on the right are the remaining consensus TEs after each step (see text for detail). Three different types of compound TEs are shown (a, b and c). Open and solid bars represent different TEs from different families. (a) One TE inserted into another. (b) Two different adjacent TEs. (c) Two adjacent copies from the same TE family.

Classification of TEs discovered by MITE-Hunter. In addition to 58 false positives, I were unable to classify 15 TEs into superfamilies. Although these sequences appeared to be TEs (based on their MSA files), their TSDs and TIRs were ambiguous because they contained too many mismatches. As such, they were judged to be unknowns.

The remaining 627 TEs were confirmed to be “cut-and-paste” class 2 TEs and were classified into previously described superfamilies. However, during the classification process I found that several families contain TEs belonging to more than one superfamily. By comparing their sequences, I discovered that this problem was caused by 14 compound TEs that were formed by the insertion of one superfamily member into another [**Figure 3.2(a)**]. Because TEs were grouped into families based on their similarity, these 14 compound TEs drag TEs from different superfamilies together. In addition, I identified another 12 compound TEs that were formed by the fusion of two

TEs from the same superfamily [**Figure 3.2, (b) and (c)**]. These 26 compound TEs have low full-length copy number in the genome and were excluded from the following analysis. Thus 601 TE consensus sequences remained.

Manual curation reveals that some TE consensus sequences in the MITE-Hunter output miss or have additional sequences at their ends. This problem is caused by the existence of false TIR and TSD structures near the authentic ones. The missing or additional sequences are mostly short and can be manually identified after locating the real TIRs and TSDs in the MSA files. After correcting the consensus sequences of the remaining 601 DNA TEs (by adding or trimming the missing or additional sequence), the similarity between some TE sequences satisfies the grouping criteria in Step III (**Figure 3.1C**). As such I ran the programs in Step III and Step V of MITE-Hunter and got the final dataset composed of 551 TE consensus sequences grouped into 401 families. Of these, 97 *Tc1/Mariner* TEs are grouped into 86 families, 146 *PIF/Harbingers* into 104 families, 123 *hATs* into 95 families, 173 *Mutators* into 110 families and 12 *CACTAs* into 6 families.

Identification of MITEs from MITE-Hunter output. To identify and characterize MITEs from MITE-Hunter output, I performed a RepeatMasker search of the rice genomic database using the curated 551 TE sequences as the query. From the RepeatMasker output, I counted the copy number of each consensus TE (data not shown). I defined a MITE as a class 2 nonautonomous TE of less than 800 bp and with at least 100 full-length copies in the genome. To include putative newborn MTIEs that don’t have such a high copy number, elements that are less than 100 full-length copies

but have at least 10 nearly identical full-length copies (identity >= 99%) were also considered as MITEs. Based on these criteria, I identified 132 rice MITEs from MITE-Hunter output, including 15 *hAT-MITEs*, 22 *Mutator-MITEs*, 50 *Stowaways* and 45 *Tourists*. No MITEs were found in the *CACTA* superfamily.

Comparison of MITE-Hunter output to Repbase data. To estimate the false negative rate of MITE-Hunter I used the rice class 2 nonautonomous elements in the Repbase as the reference dataset. Repbase was selected for this analysis because it is a collective TE database containing most, if not all, previously reported rice class 2 TEs (Jurka *et al.* 2005). However, because Repbase contains both class 1 and class 2 autonomous and nonautonomous TEs, the first step was to retrieve only rice class 2 nonautonomous elements. From these I then selected 230 elements that were less than 1.7 kb because the longest rice TE found by MITE-Hunter has 1,676 bp. The 230 elements were manually checked using the same approach that was applied to the MITE-Hunter output. Thirty-two of the 230 elements were excluded because they lack multiple full-length copies. In addition, 13 were excluded because their TIR and TSD structures could not be identified from MSA files. The remaining 185 Repbase TEs were classified into class 2 TE superfamilies. By using the same approach as used for identifying MITEs from the MITE-Hunter output, I identified 101 MITE-like elements from the 185 Repbase TEs, including 4 *hAT-MITEs*, 19 *Mutator-MITEs*, 40 *Stowaways* and 38 *Tourists*.

The false negative rates of MITE-Hunter were calculated separately for class 2 nonautonomous TEs and MITEs as follows. First, I used the curated 551 class 2

nonautonomous TEs discovered by MITE-Hunter as the query to mask the Repbase dataset using RepeatMasker. On average, 84.9% of the sequences in the Repbase dataset were masked [**Table 3.1, column (a)**]. Using a similar approach, 97.6% of MITE sequences in Repbase were masked by the TEs in the MITE-Hunter output [**Table 3.1, column (b)**]. Thus the false negative rate of MITE-Hunter is 15.1% for class 2 nonautonomous TEs and 2.4% for MITEs. MITE-Hunter failed to identify only two *Tourist* MITEs (*OSTE23* and *ID-4*) that were in Repbase. In contrast, using the data of the Repbase as the libraries, 47.9% of class 2 nonautonomous TEs and 83.4% of MITEs in the MITE-Hunter output were masked (**Table 3.1**, the last two columns). Sixteen MITEs discovered by MITE-Hunter were not found in Repbase including one *Tourist*, eleven *hAT-MITEs* and four *Mutator-MITEs*.

Table 3.1. Comparison between MITE-Hunter output and TEs in Repbase

Superfamily	**Repbase data masked by MITE-Hunter output (%)**		**MITE-Hunter output masked by Repbase data (%)**	
	All [(a)]	**MITEs only [(b)]**	**All [(c)]**	**MITEs only [(d)]**
Tc1/Mariner	93.3	100.0	72.5	99.9
PIF/Harbinger	83.8	94.6	53.1	93.0
hAT	85.8	100.0	25.6	28.4
Mutator	81.0	99.3	49.5	80.0
CACTA	88.2	-	81.7	-
Together	84.9	97.6	47.9	83.4

[(a)] 185 rice class 2 nonautonomous TEs that are less than 1.7 kp in Repbase

[(b)] 101 MITEs identified and isolated from the dataset [(a)]

[c] 551 class 2 nonautonomous TE consensus sequences curated from the MITE-Hunter output

[d] 132 MITEs identified and isolated from the dataset [c]

Evaluation of FINDMITE and MUST. I tested the ability of two previously published MITE finding programs, FINDMITE and MUST, to discover MITEs in the rice genomic dataset using default parameters. Importantly, when I attempted to use the entire genomic sequence (about 372.8 Mb) as the input data, both FINDMITE and MUST reported errors and quit. As such I applied FINDMITE and MUST to a much smaller dataset, rice chromosome 12 (about 28.2 Mb). MUST completed the task in about five hours and thirty minutes and generated 5,485 putative TE sequences. Because FINDMITE requires users to define the TSD sequence and length, I chose "TA", which is the TSD sequence of *Stowaway* MITEs. FINDMITE finished in less than 1 minute and generated 10,864 putative *Stowaways*. To calculate the false positive rate, I randomly sampled 100 TE sequences from the outputs of FINDMITE and MUST, respectively, and checked them using the same approach as was used for evaluating MITE-Hunter. With only 15 and 14 validated TEs for FINDMITE and MUST, respectively, both programs have a false positive rate over 80%. To perform an impartial comparison, I also applied MITE-Hunter to the rice chromosome 12 dataset. Using default parameters, MITE-Hunter finished in one hour and forty minutes and generated 114 TE consensus sequences that were grouped into 88 families. Through manual curation, five TEs were identified as false positives resulting in a false positive

rate of 4.4%. Because the input data is a small subset of the rice genome, I did not compare the results of FINDMITE and MUST to the Repbase data to calculate the false negative rate.

Discussion

A necessary prerequisite for the comprehensive analysis of MITEs is their identification in newly sequenced genomes. Two programs were previously developed for this purpose, FINDMITE and MUST. However, as demonstrated in this study, both FINDMITE and MUST have very high false positive rates (about 85%) and cannot efficiently utilize whole genomic datasets like that from rice. To remedy this situation, I developed MITE-Hunter, which is a structure-based program pipeline that can efficiently identify TEs that have TIR and TSD structures from whole genome datasets. Important features of MITE-Hunter are discussed below.

MITE-Hunter has an efficient approach to reduce high false positive rate, which is the main limitation of currently available MITE discovery programs. The vast majority of rice genomic sequences with TIR-like and TSD-like structures are not class 2 TEs. MITE-Hunter has two modules to filter false positives, that both exploit the principle that homologs of a true TE only share sequence similarity within the terminal structures. The main difference between the two modules is that one detects sequence similarity through the PSA approach while the other uses the MSA approach. The MSA-based module is more powerful at identifying false positives but it is slower than the PSA-based module. To achieve both high speed and high sensitivity, the PSA-based module is first performed in Step II to filter most false positives while the MSA-based module is performed in step IV to filter the remaining false positives. Because MITE-Hunter has such a system to identify and filter artificial TE candidates, the false positive rate of

MITE-Hunter (4.4% - 8.3%) is ten times lower than either FINDMITE (85%) or MUST (86%).

MITE-Hunter is competent at discovering class 2 nonautonomous TEs especially MITEs. In our test, MITE-Hunter rediscovered most of the known rice class 2 nonautonomous TEs (85%) and almost all MITEs (97.6%) in Repbase [**Table 3.1**, columns (a) and (b)]. Only two MITEs (*OSTE23 and ID-4*) in Repbase were missed by MITE-Hunter. *OSTE23* is a very old MITE family and its TIR and TSD structures are difficult to detect even by manual examination of the MSA file. *ID-4* has two mismatches in the TIRs that were not identified in Step I of MITE-Hunter.

Compared to other MITE discovery programs, the MITE-Hunter output is much easier to curate manually. First, the number of TEs in the MITE-Hunter output is very small because MITE-Hunter generates consensus sequences that best represent the whole TE dataset of the genome being analyzed. As shown in the results section, MITE Hunter generated 700 consensus TEs from the entire rice genomic dataset. In contrast, FINDMITE generated about ten thousand putative *Stowaway* MITEs using only the smallest rice chromosome (#12) as the input dataset. Using the same dataset MUST generated about five thousand elements. Second, for each TE sequence in its output, MITE-Hunter generates a MSA file and predicts TSDs, which are useful for both TE validation and classification. The validity of each TE discovered by MITE-Hunter can be determined by identifying TIRs and TSDs from the MSA file by manual inspection. Finally, in the output of MITE-Hunter, identified TEs are automatically grouped into families based on the sequence similarity, which further help manual curation by users.

These features are of value to all users, especially those who need a TE dataset that is 100% accurate and is classified into superfamilies

In summary, MITE-Hunter is the first program to efficiently and accurately identify MITEs from whole genome sequence. Whereas the rice class 2 nonautonomous TEs in Repbase were the products of many studies, MITE-Hunter was able to find virtually all the MITEs in a relatively short time frame and to do so accurately. Finally, the MITE-Hunter output is easy to curate as it contains highly condensed TE consensus sequences that are grouped into families. The validity of a TE discovered by MITE-Hunter can be quickly judged from the automatically generated MSA file, which is, to our knowledge, a unique feature of MITE-Hunter.

Acknowledgements

I thank Yaowu Yuan for valuable discussions of both of the programs and the manuscript. I thank Hao Wang for installing and running MUST. This work was supported by the NSF plant genome grant 0607123.

References

AFA, S., R, H., and P, G. 2004. RepeatMasker. *http://www.repeatmasker.org*.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Bergman, C.M. and Quesneville, H. 2007. Discovering and detecting transposable elements in genome sequences. *Brief Bioinform* **8:** 382-392.

Bureau, T.E., Ronald, P.C., and Wessler, S.R. 1996. A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. *Proc Natl Acad Sci U S A* **93:** 8524-8529.

Bureau, T.E. and Wessler, S.R. 1992. Tourist: a large family of small inverted repeat elements frequently associated with maize genes. *Plant Cell* **4:** 1283-1294.

Bureau, T.E. and Wessler, S.R. 1994. Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell* **6:** 907-916.

Chen, Y., Zhou, F., Li, G., and Xu, Y. 2009. MUST: a system for identification of miniature inverted-repeat transposable elements and applications to Anabaena variabilis and Haloquadratum walsbyi. *Gene* **436:** 1-7.

Dooner, H.K. and Weil, C.F. 2007. Give-and-take: interactions between DNA transposons and their host plant genomes. *Curr Opin Genet Dev* **17:** 486-492.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32:** 1792-1797.

Feschotte, C., Jiang, N., and Wessler, S.R. 2002. Plant transposable elements: where genetics meets genomics. *Nat Rev Genet* **3:** 329-341.

Feschotte, C. and Pritham, E.J. 2007. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* **41:** 331-368.

Feschotte, C., Swamy, L., and Wessler, S.R. 2003. Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). *Genetics* **163:** 747-758.

Han, Y., Burnette, J.M., 3rd, and Wessler, S.R. 2009. TARGeT: a web-based pipeline for retrieving and characterizing gene and transposable element families from genomic sequences. *Nucleic Acids Res* **37:** e78.

Jiang, N., Bao, Z., Zhang, X., Hirochika, H., Eddy, S.R., McCouch, S.R., and Wessler, S.R. 2003. An active DNA transposon family in rice. *Nature* **421:** 163-167.

Jiang, N., Feschotte, C., Zhang, X., and Wessler, S.R. 2004. Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). *Curr Opin Plant Biol* **7:** 115-119.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110:** 462-467.

Kirkness, E.F., Bafna, V., Halpern, A.L., Levy, S., Remington, K., Rusch, D.B., Delcher, A.L., Pop, M., Wang, W., Fraser, C.M. et al. 2003. The dog genome: survey sequencing and comparative analysis. *Science* **301:** 1898-1903.

Kuang, H., Padmanabhan, C., Li, F., Kamei, A., Bhaskar, P.B., Ouyang, S., Jiang, J., Buell, C.R., and Baker, B. 2009. Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: new functional implications for MITEs. *Genome Res* **19:** 42-56.

Lander, E.S. Linton, L.M. Birren, B. Nusbaum, C. Zody, M.C. Baldwin, J. Devon, K. Dewar, K. Doyle, M. FitzHugh, W. et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lerat, E. 2009. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity* **104:** 520-533.

Moreno-Vazquez, S., Ning, J., and Meyers, B.C. 2005. hATpin, a family of MITE-like hAT mobile elements conserved in diverse plant species that forms highly stable secondary structures. *Plant Mol Biol* **58:** 869-886.

Naito, K., Cho, E., Yang, G., Campbell, M.A., Yano, K., Okumoto, Y., Tanisaka, T., and Wessler, S.R. 2006. Dramatic amplification of a rice transposable element during recent domestication. *Proc Natl Acad Sci U S A* **103:** 17620-17625.

Oki, N., Yano, K., Okumoto, Y., Tsukiyama, T., Teraishi, M., and Tanisaka, T. 2008. A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, Oryza sativa ssp. japonica. *Genes Genet Syst* **83:** 321-329.

Osborne, P.W., Luke, G.N., Holland, P.W., and Ferrier, D.E. 2006. Identification and characterisation of five novel miniature inverted-repeat transposable elements (MITEs) in amphioxus (Branchiostoma floridae). *Int J Biol Sci* **2:** 54-60.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A. et al. 2009. The Sorghum bicolor genome and the diversification of grasses. *Nature* **457:** 551-556.

Santiago, N., Herraiz, C., Goni, J.R., Messeguer, X., and Casacuberta, J.M. 2002. Genome-wide analysis of the Emigrant family of MITEs of Arabidopsis thaliana. *Mol Biol Evol* **19:** 2285-2293.

Schnable, P.S. Ware, D. Fulton, R.S. Stein, J.C. Wei, F. Pasternak, S. Liang, C. Zhang, J. Fulton, L. Graves, T.A. et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326:** 1112-1115.

Smit, A.F. and Riggs, A.D. 1996. Tiggers and DNA transposon fossils in the human genome. *Proc Natl Acad Sci U S A* **93:** 1443-1448.

Surya, S., Susan, B., Zenaida, V.M., and Daniel, G.P. 2008. Computational approaches and tools used in identification of dispersed repetitive DNA sequences. *Tropical Plant Biol.***:** 85-96.

Tanaka, T., Antonio, B.A., Kikuchi, S., Matsumoto, T., Nagamura, Y., Numa, H., Sakai, H., Wu, J., Itoh, T., Sasaki, T. et al. 2008. The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res* **36:** D1028-1033.

Tu, Z. 2001. Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, Anopheles gambiae. *Proc Natl Acad Sci U S A* **98:** 1699-1704.

Waterston, R.H. Lindblad-Toh, K. Birney, E. Rogers, J. Abril, J.F. Agarwal, P. Agarwala, R. Ainscough, R. Alexandersson, M. An, P. et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O. et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8:** 973-982.

Yang, G., Nagel, D.H., Feschotte, C., Hancock, C.N., and Wessler, S.R. 2009. Tuned for transposition: molecular determinants underlying the hyperactivity of a Stowaway MITE. *Science* **325:** 1391-1394.

Zhang, X., Feschotte, C., Zhang, Q., Jiang, N., Eggleston, W.B., and Wessler, S.R. 2001. P instability factor: an active maize transposon system associated with the amplification of Tourist-like MITEs and a new superfamily of transposases. *Proc Natl Acad Sci U S A* **98:** 12572-12577.

Chapter 4

Characterization of the abundance, variation, distribution and impact on gene expression of class 2 nonautonomous transposable elements in grass genomes

---

Yujun Han and Susan R. Wessler. 2010.

To be submitted to Proceedings of the National Academy of Sciences.

Abstract

Class 2 nonautonomous transposable elements (TEs), especially miniature inverted-repeat transposable elements (MITEs), are present at very high copy numbers in most sequenced genomes. Among cereal grasses, they are the predominant TEs that are in or near host genes. The availability of four sequenced grass genomes provides a unique opportunity to characterize TEs and determine their genome wide association with grass genes. To this end I used two newly developed programs, TARGeT and MITE-Hunter, to annotate class 2 TEs in the genomes of four members of the grass family: *Brachypodium distachyon*, *Oryza sativa*, *Sorghum bicolor* and *Zea mays* ssp. *mays* L. Unlike previously described programs, the pipeline used in this study facilitated the accurate and efficient identification and classification of class 2 nonautonomous TEs and MITEs from whole genome sequences. Three features of class 2 nonautonomous TEs were revealed in this study. First, distinctive patterns of TE length and copy number were found for each superfamily. This result provides new insights into the origin and amplification of MITEs. Second, the association of class 2 nonautonomous TEs and MITEs with genes is described at high resolution in grass genomes. Specifically, TEs from all superfamilies were found to be most abundant in the 5' gene flanking regions with the highest density found within 1 kb of the transcription start site. Finally, attempts were made to correlate the presence or absence of TEs in the 5' flanking regions of genes with different levels of gene expression. Interestingly, this analysis revealed that highly expressed genes were more likely to contain TEs in their 5' flanking region than genes with lower levels of expression. Taken together, this study provides the most

comprehensive analysis to date of the association and potential impact of class 2 nonautonomous TEs and MITEs with grass genes.

Introduction

Transposable elements (TEs) are mobile DNA fragments that can move from one genomic locus to another, by a process called transposition. TEs are divided into class 1 (RNA) TEs and class 2 (DNA) TEs based on whether the intermediate during transposition is RNA or DNA. TEs in each class are further divided into superfamilies and families based on their sequence identity (Wicker *et al.* 2007). To date, more than ten superfamilies have been found among class 2 TEs and six of these have been reported in plants, including *Helitron*, *CACTA*, *Tc1/Mariner*, *hAT*, *Mutator* and *PIF/Harbinger* (Feschotte and Pritham 2007; Wicker *et al.* 2007). In addition to sequence relatedness, each superfamily has unique structural features that can be used in their classification. All superfamily members except *Helitron*s transpose by a "cut-and-paste" mechanism and have terminal inverted repeats (TIRs) and generate target site duplications (TSDs) upon insertion. All superfamilies contain autonomous and nonautonomous family members where the autonomous TEs encode the transposase necessary for transposition. In addition to moving themselves, autonomous TEs are able to mobilize nonautonomous elements, which lack functional transposase either because nonautonomous TEs contain a mutated coding region or no coding region at all.

Although the first TE was discovered more than half of a century ago (McClintock 1947; McClintock 1948), it has only recently been shown that TEs make up the majority of the genomes of most higher eukaryotes. For example, TEs occupy at least half of the genomic sequences of human (*Homo sapiens*) and about 85% of maize (*Zea mays* ssp.

mays L.) (Lander *et al.* 2001; Schnable *et al.* 2009). In many plant genomes the predominant TEs are class 1 LTR retrotransposons and class 2 miniature inverted-repeat transposable elements (MITEs). MITEs, which are the focus of this study, are a special type of class 2 nonautonomous element of short length (most < 500 bp) and genome copy numbers that can attain hundreds or thousands (Bureau and Wessler 1992; Bureau and Wessler 1994). Although some MITEs have been discovered in the *hAT* and *Mutator* superfamilies (Kuang *et al.* 2009; Moreno-Vazquez *et al.* 2005), most MITEs are either *Stowaway* or *Tourist* elements that belong to the *Tc1/Mariner* and *PIF/Harbinger* superfamilies, respectively (Feschotte *et al.* 2003; Jiang *et al.* 2003; Yang *et al.* 2009; Zhang *et al.* 2004).

Like other class 2 TEs, MITEs prefer to insert near or within host genes (Naito *et al.* 2006; Naito *et al.* 2009; Spradling *et al.* 1995). Given their high copy numbers, this tendency has evolutionary implications especially in light of a recent study of new insertions of the rice MITE *mPing*, where it was shown that *mPing* insertions in promoter regions preferentially up regulate nearby gene transcription (Naito *et al.* 2009).
The availability of increasing numbers of sequenced genomes provides an opportunity to understand MITE origin, spread and their impact on the host. However, until very recently the study of MITEs has been hindered by an inability to identify them in genome sequence. As described in the previous chapter, I developed a MITE discovery program pipeline called MITE-Hunter that efficiently and accurately finds MITEs in genome sequence. In this chapter, I used MITE-Hunter to discover class 2 nonautonomous TEs from four available grass genomic datasets, *Brachypodium*

distachyon, rice (*Oryza sativa*), sorghum (*Sorghum bicolor*) and maize (*Zea mays* ssp. *mays* L.) (2005; Paterson *et al.* 2009; Schnable *et al.* 2009). Comparisons of MITEs and their genomic distribution in these four species have led to three important findings. First, MITEs from different superfamilies have distinct features with regard to size and copy number. Second, in all species examined, MITEs are preferentially associated with the 5' flanking regions of genes. Finally, there is a significant positive correlation between the presence of a MITE near a gene and the expression level of that gene.

Materials and Methods

Datasets. The following versions of genomic sequences and corresponding gene annotation files were used in this study: Bd21 and Bradi_1.0.gff2 of *B. distachyon* (http://www.brachypodium.org/) , IRGSP build 5 (http://rgp.dna.affrc.go.jp/IRGSP/) and RAP3.gff3 (http://rapdb.dna.affrc.go.jp/) of rice (2005; Tanaka *et al.* 2008), Sorbi1 and Sbi1_4.gff of sorghum (http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html) (Paterson *et al.* 2009), and 4a53 and 4a.53_FGS.gff of maize (http://www.maizesequence.org/index.html) (Schnable *et al.* 2009). The rice microarray dataset, GEO series GSE15021, was downloaded from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) (Barrett and Edgar 2006) for the gene expression analysis.

Discovery of TEs. Class 2 TEs were discovered from the above genomic sequences as follows. Autonomous elements were identified by TARGeT (Han *et al.* 2009), using sequences from the conserved regions of transposase proteins as the query. Class 2 TE transposase sequences were downloaded from Repbase (Jurka *et al.* 2005). Nonautonomous TEs were discovered by MITE-Hunter (unpublished), which specifically identifies class 2 nonautonomous TEs and provides consensus sequences as the output. TARGeT and MITE-Hunter were run on a Linux cluster with default parameters.

Curation of TEs discovered by MITE-Hunter. Outputs of MITE-Hunter were manually verified to eliminate false positives together with unknown and compound TEs. TEs were classified into superfamilies based on the TIR and TSD sequences using the

following rules. TEs that have TIRs starting with CACTA/G and have 3 bp TSDs are identified as *CACTA* elements. TEs with 8 bp TSDs are identified as *hAT* elements. TEs with 9 or 10 bp TSDs are identified as *Mutator* elements. TEs with 2 bp TSDs that are "TA" are identified as *Tc1/Mariner* elements. TEs with 3 bp TSDs that are TAA or TTA are identified as *PIF/Harbinger* elements.

Calculation the total length and copy numbers of TEs. Copy numbers of class 2 autonomous TEs were determined from the TARGeT outputs. Total length and copy numbers of class 2 nonautonomous TEs were calculated from RepeatMasker outputs as follows. First, curated class 2 nonautonomous TE sequences were used as library files for RepeatMaker (version 3.26, http://www.repeatmasker.org) to mask the genomic sequences with "–nolow" and "–no_is" parameters. Second, because the positions of some TE copies overlap in the RepeatMasker output, to avoid counting the same TE copy twice, the RepeatMasker output was curated by a Perl script. In the RepeatMasker output, if two consecutive TE copies overlap, the start position of the second one will be changed to be right after the stop position of the first one. If a short TE copy is within a longer one, the short one will be filtered.

TE copies and MITEs were determined from the RepeatMasker output as follows. Because one TE copy may degenerate into several fragments in the genome, the number of total fragments is much higher than the actual total copy number of TEs. I counted the TE copy number using an approach that is similar to one introduced previously (Schnable *et al.* 2009). When compared to the query TE sequence, if a TE match identified by RepeatMasker misses less than 20 bp on either side it was

considered as a full-length copy (FC). If a TE match has only one end that miss less than 20 bp, it is counted as a copy with a missing end (EC). Other TE matches in the RepeatMasker output are considered as fragments that were not counted into copy numbers. A TE is considered to be a MITE if it is less than 800 bp and has at least 100 FCs, or 10 FCs with an identity >= 99%.

Analysis of the association of class 2 TEs and genes. Several Perl scripts were developed to retrieve the genomic positions of TE copies and genes and to draw figures showing the distribution of TE proportions within and around genes. The positions of TE copies were retrieved from RepeatMasker outputs. The positions of genes were retrieved from gene annotation files as follows. First, the gene annotation files were checked to filter genes that are inside other genes and genes with obvious incorrect annotation. Second, for each gene, positions were retrieved from different compartments, including 5' flanking region (5FR), 5' UTR exons (5UE), 5' UTR introns (5UI), exons (E), introns (I), 3' UTR exons (3UE), 3'UTR introns (3UI) and 3' flanking region (3FR). The default length of the flanking region is 10 kb. If the distance between two genes is shorter than 20 kb, the length of the flanking regions between the genes will be half of the distance between them. If a gene has multiple annotated modules, a combination gene module was generated by merging multiple gene modules together. Proportions of TE sequences within and around genes were calculated as follows. Except for the 5FR and the 3FR, for each type of gene compartment, an average percentage was calculated by dividing the total length of TEs within the gene compartment by the total length of the gene compartment. Percentages of TEs were

calculated for each nucleotide position within the 5FR and the 3FR and used to generate a graphical output.

Analysis of gene expression. Rice genes were divided equally into three groups based on their expression levels in microarray results stored in GSE15021 (Naito *et al.* 2009). For each group, proportions of TEs within and around genes were calculated using the approach described above. To determine the statistical significance, pairwise comparisons were conducted on proportions of TEs between groups using Wilcoxon Two-Sample Test in SAS 9.1. The level of significance was adjusted to be .003 using Bonferroni correction.

Results

Abundance of class 2 TEs in grass genomes. Class 2 autonomous and nonautonomous TEs were annotated from the genomic sequences of *B. distachyon*, rice, sorghum and maize using TARGeT, MITE-Hunter and RepeatMasker as described in the *Materials and Methods*. The results are summarized below where emphasis is put on the annotation of nonautonomous TEs (see details in **Table 4.1**).

Table 4.1 Class 2 TEs in B. *distachyon*, rice, sorghum and maize

Species	Superfamilies	Autonomous copies [b]	Nonautonomous [a]					
				No. of consensus [c]	Full-length copies	Copies that have only one end	Total length (Mb)	% of genome
Brachypodium (270.8 Mb)	*CACTA*	633	MITEs [d]	0	0	0	0.00	0.00
			Others [e]	2	97	113	0.08	0.03
	hAT	175	MITEs	4	673	587	0.20	0.07
			Others	70	1,251	4,347	1.12	0.41
	Mutator	515	MITEs	4	1,256	1,790	0.06	0.02
			Others	69	1,112	8,090	2.10	0.78
	PIF/Harbinger	306	MITEs	11	2,331	3,431	1.11	0.41
			Others	52	1,730	5,990	1.50	0.55
	Tc1/Mariner	38	MITEs	33	22,231	15,743	4.12	1.52
			Others	44	1,075	1,989	0.36	0.13
Subtotal		**1,667**		**289**	**31,756**	**42,080**	**10.65**	**3.93**
Rice (372.8 Mb)	*CACTA*	1,740	MITEs	0	0	0	0.00	0.00
			Others	12	247	6,535	3.11	0.83
	hAT	279	MITEs	12	1,875	1,857	0.06	0.02
			Others	111	2,883	9,779	3.35	0.90
	Mutator	705	MITEs	21	3,945	7,070	2.52	0.68
			Others	152	3,543	36,330	9.37	2.51
	PIF/Harbinger	215	MITEs	44	26,691	23,693	10.15	2.72
			Others	102	2,543	8,734	2.61	0.70

	Tc1/Mariner	45	MITEs	50	28,520	23,344	8.65	2.32
			Others	47	1,837	3,537	0.56	0.15
Subtotal		**2,984**		**551**	**72,084**	**120,879**	**40.38**	**10.83**
Sorghum (697.6 Mb)	*CACTA*	3,104	MITEs	1	401	190	0.09	0.01
			Others	9	203	5,282	3.73	0.53
	hAT	531	MITEs	10	2,289	1,832	0.79	0.11
			Others	149	3,221	10,411	3.73	0.53
	Mutator	419	MITEs	4	783	1,478	1.36	0.19
			Others	54	924	6,289	2.55	0.37
	PIF/Harbinger	388	MITEs	43	51,513	33,045	17.36	2.49
			Others	184	3,825	15,258	4.56	0.65
	Tc1/Mariner	57	MITEs	25	13,145	14,701	5.20	0.75
			Others	44	782	2,427	0.56	0.08
Subtotal		**4,499**		**523**	**77,086**	**90,913**	**39.93**	**5.72**
Maize (2061.0 Mb)	*CACTA*	4,324	MITEs	2	308	6,056	2.34	0.11
			Others	14	338	7,741	3.90	0.19
	hAT	1,569	MITEs	46	11,187	14,511	7.99	0.39
			Others	263	5,513	29,879	12.61	0.61
	Mutator	1,157	MITEs	6	1,662	786	0.38	0.02
			Others	81	1,937	3,855	2.44	0.12
	PIF/Harbinger	1,523	MITEs	137	46,669	43,453	15.51	0.75
			Others	277	8,300	20,597	6.28	0.30
	Tc1/Mariner	47	MITEs	27	10,595	10,095	2.74	0.13
			Others	22	750	1,366	0.35	0.02

Subtotal		**8,620**		**875**	**87,259**	**138,339**	**54.54**	**2.65**

[a] Only elements that are less than 2 kb are included.

[b] These elements are copies of conserved regions of transposase genes and may not have full length copies.

[c] These are curated class 2 nonautonomous TE consensus sequences generated by MITE-Hunter.

[d] We defined a MITE as a class 2 nonautonomous TE that is shorter than 800 bp with at least 100 full-length copies or 10 almost identical full-length copies (identity >= 99%).

[e] These are “traditional” low copy nonautonomous TEs that are not MITEs as defined above [d].

Although the genome size among these four species is dramatically different, the total length of class 2 nonautonomous TEs is similar except for *B. distachyon*. For example, while the genome size of sorghum is almost twice as large as that of rice (~ 700 Mb vs. ~ 370 Mb), the total TE length is virtually the same (~40 Mb for both). Similarly, while the maize genome is almost 6 fold larger than the rice genome, it only has ~ 35% more class 2 nonautonomous TEs (~54Mb vs. 40Mb).

My data also indicates that there are significant differences in copy number among superfamily members and species. Of the plants examined, *Tourist* MITEs are most numerous in both sorghum [~52,000 full-length copies (FCs)] and maize (~47,000 FCs) and are also abundant in rice (~27,000 FCs). Although the number of autonomous *Tc1/Mariner* elements is the smallest (~ 50 copies in each genome), *Stowaway* MITEs have the highest copy number in the *B. distachyon* (~ 22,000 FCs) and rice (~29,000 FCs) genome. In contrast, the *CACTA* superfamily has both the highest number of autonomous copies and the lowest number of *CACTA-like* MITEs in all four grass species.

To evaluate the quality of the above annotation results, I compared them with results of other studies (2010; Oki *et al.* 2008; Schnable *et al.* 2009). Because the criteria used to calculate copy number is unknown from the publications of these studies, the percentage of total TE length was used for the comparison. In summary, I found 8% to 32% more MITE sequences in rice, about two times more in both *B. distachyon* and sorghum, and about four times more in maize (**Table 4.2**). Because the rules to define MITEs are also not presented in these publications, I am not sure

whether there are some elements considered as MITEs in my results but not in other studies, because they may use stricter criteria to define a MITE. To this end, I did another comparison, in which the total length of MITEs in my annotation result was used to compare the total length of all of the elements in each superfamily, including both autonomous and nonautonomous elements, in these publications. Even under such an unfair condition, except for *B. distachyon* where I found 23% more MITE sequences, there are no significant changes of the ratio in rice, sorghum and maize.

Table 4.2 Comparison between MITEs in my annotation results with data in previous publications

MITEs	Proportion of the genome (%) [(a)]				
	B. distachyon	Rice		Sorghum	Maize
Stowaway	0.88/1.52	1.60/2.32	1.74/2.32	0.19/0.75	0.03/0.13
Tourist	0.18/0.41	1.50/2.72	1.50/2.72	0.94/2.49	0.08/0.75
Others	na/0.09	0.79/0.7	2.00/0.70	0.61/0.31	0.21/0.52
Total	1.06/2.02	3.89/5.72	5.24/5.72	1.74/3.55	0.32/1.40

(a) There are two numbers in each column. The first number is retrieved from results of previous publications. The second number is from the results of this study.

Superfamilies produce nonautonomous elements with distinguishing features.

The availability of hundreds of thousands of nonautonomous class 2 TEs presented an opportunity to see if interspecies comparisons would reveal any distinguishing features of superfamily members. To this end, graphs were generated to compare element copy number and length for different superfamilies in each species (**Figure 4.1**). *CACTA* TEs

were not included because only a few *CACTA* consensus sequences were generated for each species.

From **Figure 4.1** it can be seen that each superfamily has unique features that are largely conserved in the fours species. To make the comparisons easier to visualize, the same axes values are used throughout. Among all superfamilies, the nonautonomous *Tc1*/*Mariner* elements have the most compact structure (most < 400 bp) and many of them (*Stowaway* MITEs) have high copy numbers (**Figure 4.1 A, E, I and M**). For example, the *Stowaway* element with the highest copy number in maize (~ 2,100 FCs) is only 80 bp (**Figure 4.1 M**). Compared to *Tc1/Mariner* elements, *PIF/Harbinger* elements also have high copy numbers and compact structures but their average length is about 100 bp longer (**Figure 4.1 B, F, J and N**). Unlike elements in *Tc1/Mariner* and *PIF/Harbinger* superfamilies, few *hAT* and *Mutator* elements have more than one thousand copies. Except for those in maize, *hAT* nonautonomous elements appear to cluster into two groups. In one group, elements cluster around 200 bp while elements in the other group are distributed over a wider range from 400 bp to 800 bp (**Figure 4.1 C, G, K and O**). *Mutator* elements have the broadest length distribution and no significant patterns can be recognized in any species examined (**Figure 4.1 D, H, L and P**).

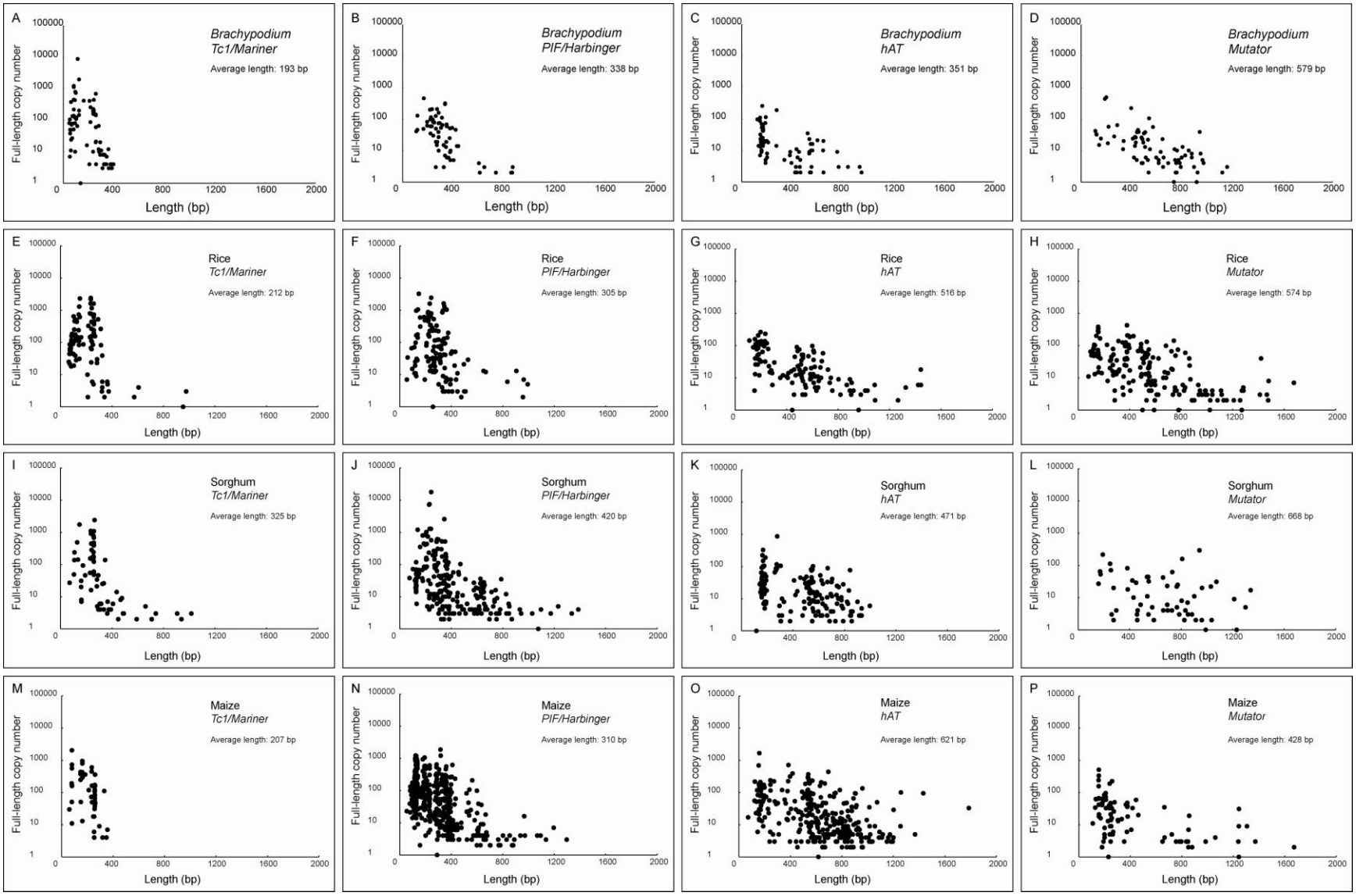

A
Brachypodium
Tc1/Mariner
Average length: 193 bp
B
Brachypodium
PIF/Harbinger
Average length: 338 bp
C
Brachypodium
hAT
Average length: 351 bp
D
Brachypodium
Mutator
Average length: 579 bp
E
Rice
Tc1/Mariner
Average length: 212 bp
F
Rice
PIF/Harbinger
Average length: 305 bp
G
Rice
hAT
Average length: 516 bp
H
Rice
Mutator
Average length: 574 bp
I
Sorghum
Tc1/Mariner
Average length: 325 bp
J
Sorghum
PIF/Harbinger
Average length: 420 bp
K
Sorghum
hAT
Average length: 471 bp
L
Sorghum
Mutator
Average length: 668 bp
M
Maize
Tc1/Mariner
Average length: 207 bp
N
Maize
PIF/Harbinger
Average length: 310 bp
O
Maize
hAT
Average length: 621 bp
P
Maize
Mutator
Average length: 428 bp
Full-length copy number
Length (bp)

Figure 4.1. Length and copy number of *Tc1/Mariner*, *PIF/Harbinger*, *hAT* and *Mutator* nonautonomous TEs identified by MITE-Hunter in Brachypodium (A - D), rice (E - H), sorghum (I - L) and maize (M - P). The X-axis is the length of element consensus sequences in bp. The Y-axis is the full-length copy number in log10.

Distribution of class 2 TEs within and around genes. It has been noted that class 2 TEs including MITEs are preferentially located close to genes (Bureau and Wessler 1992; Dietrich *et al.* 2002; Rizzon *et al.* 2003; Spradling *et al.* 1995; Zhang *et al.* 2000). Given the large numbers of elements identified in this study, we are now able to address the question, how close is close? In this study, I calculated the proportion of TE sequences within and around genes, based on the positions of TE copies and genes in each genome. The distribution of class 2 nonautonomous TEs was precisely determined for each superfamily in four grass species (see *Materials and Methods* for details).

Although there is no significant sequence similarity among the TEs belonging to different superfamilies, the comparative data reveals striking similarity in their distribution around and within genes in all the species tested **(Figure 4.2)**. The region that has the highest density of class 2 nonautonomous TE sequences is within 1 kb upstream of the transcription start site. I calculated the TE sequences content of this region in rice. On average 20% of this region is composed with TE sequences and 53% rice genes contain TE sequences longer than 80 bp in this region. The region with the

second highest density of class 2 nonautonomous TE sequences is within 1 kb downstream of the transcription stop site. The TE density patterns in the upstream and downstream regions of genes are almost symmetric, except that the latter one has a relatively lower peak. Within genes, class 2 nonautonomous TEs are most abundant within introns. However, compared to the 1 kb regions upstream and downstream of genes, the TE proportion in introns is much lower. Finally, in all species and of all TE superfamilies, class 2 nonautonomous TEs have the lowest density in exons.

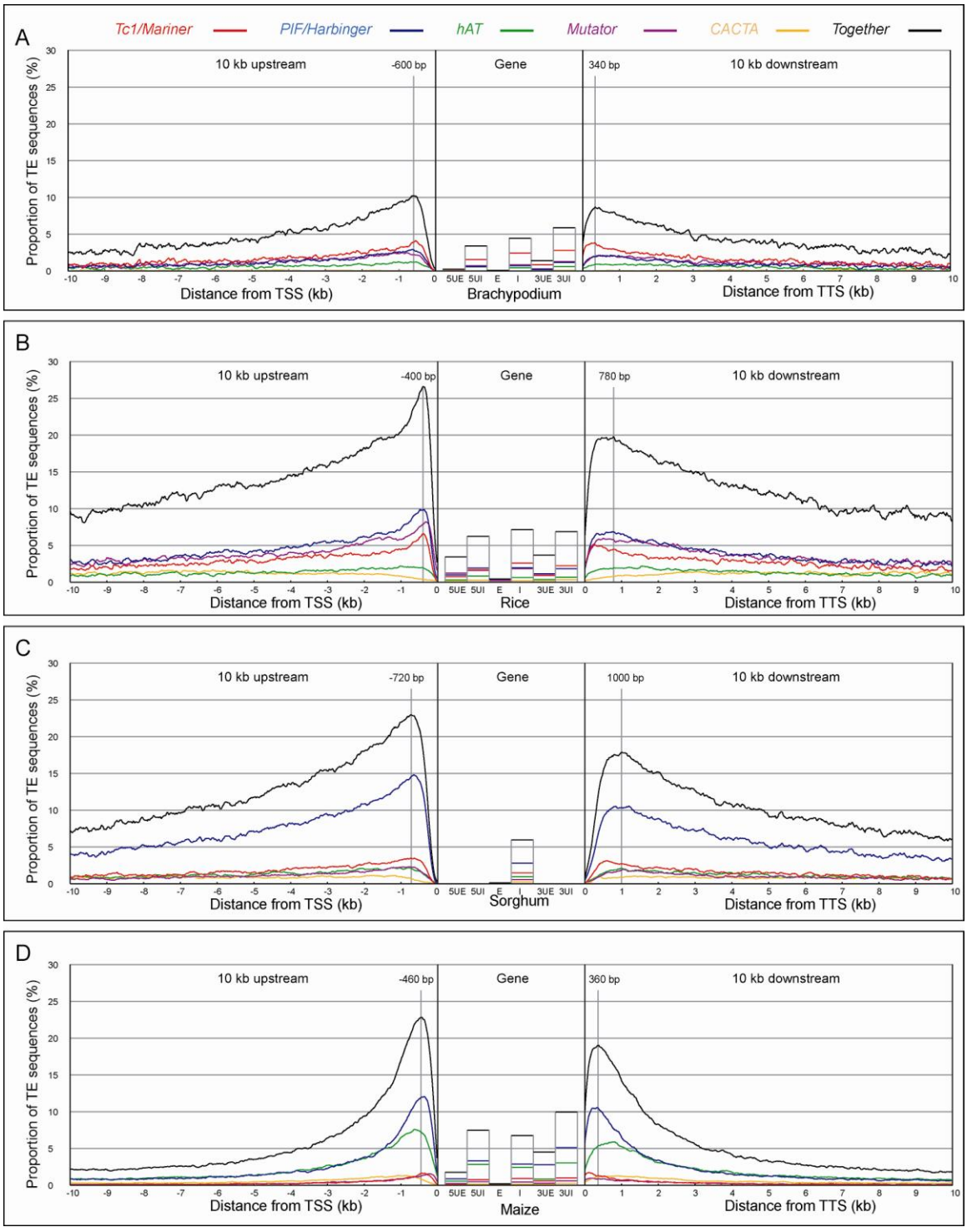


Figure 4.2. Proportion of TE sequences within and around the genes of *Brachypodium* (A), rice (B), sorghum (C) and maize (D). The Y-axis is the percentage of TE sequences

at a particular position in all genes of that genome. Black lines represent the total proportion of TEs. Colored lines represent different class 2 superfamilies: blue, green, red, purple and gold represent *PIF/Harbinger*, *hAT*, *Tc1/Mariner*, *Mutator* and *CACTA* superfamilies, respectively. The X-axis is divided into three sections. The section on the left represents the distribution of TE sequences in the 5' flanking regions of genes and shows the distance from the transcription start site (TSS). The section on the right represents the distribution of TE sequences in the 3' flanking regions of genes shows the distance from the transcription stop site (TTS). The highest TE sequence proportions in the 5' and 3' gene flanking regions are marked by vertical gray lines and the distance from the peak to the TSS or TTS is shown above the lines. The middle section shows the average proportion of TE sequences in 5' the UTR, exons (5UE), 5' UTR, introns (5UI), exons (E), introns (I), 3' UTR, exons (3UE), 3' UTR introns (3UI).

Effects on gene expression. The predominance of TEs in the promoters of the genes of these four species prompted an analysis of whether genes with TE insertions displayed distinctive expression patterns. To this end, I classified rice genes into three groups with high, medium or low expression levels and calculated class 2 TE proportions in the 1 kb flanking regions of each group of genes (see *Materials and Methods* for details). The result shows that genes with higher expression levels have more class 2 TE sequences in both their upstream (p values < 0.001) and downstream regions (p values < 0.005) (**Figure 4.3 A and B**). Such differences are more significant

when I used only *Tourist* MITEs for the analysis (p value < 0.001 in both cases) (**Figure 4.3 C and D**).

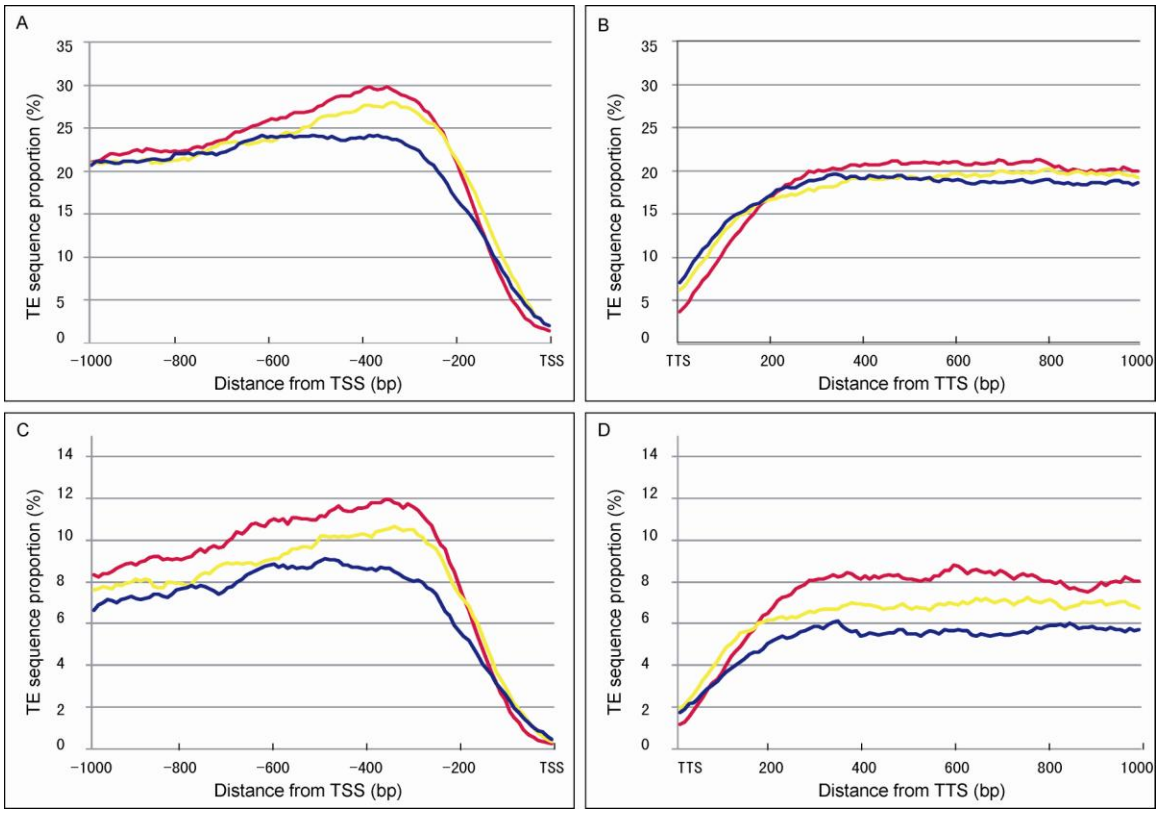


Figure 4.3. The proportions of class 2 nonautonomous TE sequences in the flanking regions of rice genes with high expression (red), medium expression (yellow) and low expression (blue). The X-axis is the distance to the TSS or the TTS. The Y-axis is the percentage of TEs. A and B) TE proportion of all class 2 TE superfamilies in the upstream and downstream of genes, respectively. C and D) TE proportion of *PIF/Harbinger* in the upstream and downstream of genes, respectively.

Discussions

Using TARGeT and MITE-Hunter, class 2 TEs were annotated and characterized in four cereal grass species, *B. distachyon*, rice, sorghum and maize. To make sure the annotation results are good for further analysis, I compared the annotation results with data in previous publications (2010; Oki *et al.* 2008; Paterson *et al.* 2009; Schnable *et al.* 2009). The comparison showed that many more MITEs were found in my annotation than in prior studies (**Table 4.2**).

The different distribution patterns of copy number and length among class 2 superfamilies provide some new insight into the mechanisms of the origination and amplification of MITEs. Compared to other superfamilies, *Tc1/Mariner* and *PIF/Harbinger* superfamilies have more MITEs and have very different distribution patterns of copy number and length (**Table 4.1**). First, the length of *Tc1/Mariner* and *PIF/Harbinger* elements is much shorter and almost no elements that are longer than 400 bp have more than 10 copies. In contrast, the length of elements in *hAT*, *Mutator* and *CACTA* superfamilies has a much broader distribution pattern and the copy numbers are not significantly increasing when the elements are getting shorter. These results indicate two features of *Tc1/Mariner* and *PIF/Harbinger* elements. First, the sequences required for efficient amplification are very short (less than about 400 bp). Second, there is a mechanism that strongly represses the amplification of long elements.

The strikingly similar distribution patterns of TE proportion in gene flanking regions lead to several obvious questions. First, does such a pattern also exist in other species than

the cereal grass family? The answer is yes because I have done a similar annotation and analysis in *Arabidopsis lyrata* and found similar TE distribution patterns (data not shown). Second, are such patterns caused by selection effects or the insertion preference of class 2 TEs? It seems impossible that they were caused by purifying selections on random TE insertions in the whole genome except for the gene flanking regions. A more plausible explanation is that class 2 nonautonomous TEs have the insertion preference in these regions. Indeed, a study of new insertion sites of an active rice *Tourist* MITE, *mPing*, strongly supports this mechanism (Naito *et al.* 2009). Why do class 2 TEs prefer inserting in a region this narrow and close to genes? One possible answer is that the nucleosome free regions flanking genes could make DNAs in this region more accessible for class 2 TEs (Lieb and Clarke 2005).

Genes with higher expression levels contain more class 2 nonautonomous TEs around them, especially *Tourist* MITEs (**Figure 4.3**). This result could be generated by two causes. First, the expression of genes makes the DNA around genes more accessible to class 2 TEs. Evidence supporting this has been reported in *D. melanogaster* (Fontanillas *et al.* 2007). Second, the TE sequence increases gene expression levels after insertion. This is supported by a recent study of *mPing* in rice, where most new insertions of *mPing* upregulate the gene expression level (Naito *et al.* 2009). If the latter model is true, because there are only ~50 copies of *mPing* in the rice genome that we analyzed, and there is no sequence similarity between *mPing* and other *Tourist* MITEs, it should be a common feature of *Tourist* MITEs to upregulate gene expression levels. However, it is also possible that these two causes work

together. My future work is to compare TE densities between germline-expressed genes with somatic-expressed genes.

In summary, using two newly developed programs, TARGeT and MITE-Hunter, class 2 TEs were annotated and characterized in four grass genomes, *B. distachyon*, rice, sorghum and maize. Conserved distribution patterns of copy number and length were identified for each superfamily, which shed new light on the mechanisms of the successful amplification of MITEs. The solid annotation results of both TEs and genes provide a good opportunity to take a close look at their position relationship in the genome. Amazingly, the region containing the highest TE proportion is within 1 kb upstream of transcription start site of genes, and further analysis reveals that genes with high expression levels have more class 2 TE sequences around them. Finally, I want to point out that this study only characterized class 2 nonautonomous TEs that exist in and around genes. Class 2 TEs can excise out, which may also lead to strong impact on gene functions, such as promoter scrambling (Kloeckener-Gruissem and Freeling 1995). Taken together, class 2 nonautonomous TEs including MITEs have a significant and complicated role in the modification of genes of grass species.

Acknowledgements

I thank Yaowu Yuan, Ken Naito, Aaron O. Richardson, Michael McKain, Xiaoyu Zhang and Jim Leebens-Mack for valuable discussions and suggestions.

References

Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* **463:** 763-768.

2005. The map-based sequence of the rice genome. *Nature* **436:** 793-800.

2010. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* **463:** 763-768.

Barrett, T. and Edgar, R. 2006. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol* **411:** 352-369.

Bureau, T.E. and Wessler, S.R. 1992. Tourist: a large family of small inverted repeat elements frequently associated with maize genes. *Plant Cell* **4:** 1283-1294.

Bureau, T.E. and Wessler, S.R. 1994. Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell* **6:** 907-916.

Dietrich, C.R., Cui, F., Packila, M.L., Li, J., Ashlock, D.A., Nikolau, B.J., and Schnable, P.S. 2002. Maize Mu transposons are targeted to the 5' untranslated region of the gl8 gene and sequences flanking Mu target-site duplications exhibit nonrandom nucleotide composition throughout the genome. *Genetics* **160:** 697-716.

Feschotte, C. and Pritham, E.J. 2007. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* **41:** 331-368.

Feschotte, C., Swamy, L., and Wessler, S.R. 2003. Genome-wide analysis of mariner-like transposable elements in rice reveals complex relationships with stowaway miniature inverted repeat transposable elements (MITEs). *Genetics* **163:** 747-758.

Fontanillas, P., Hartl, D.L., and Reuter, M. 2007. Genome organization and gene expression shape the transposable element distribution in the Drosophila melanogaster euchromatin. *PLoS Genet* **3:** e210.

Han, Y., Burnette, J.M., 3rd, and Wessler, S.R. 2009. TARGeT: a web-based pipeline for retrieving and characterizing gene and transposable element families from genomic sequences. *Nucleic Acids Res* **37:** e78.

Jiang, N., Bao, Z., Zhang, X., Hirochika, H., Eddy, S.R., McCouch, S.R., and Wessler, S.R. 2003. An active DNA transposon family in rice. *Nature* **421:** 163-167.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110:** 462-467.

Kloeckener-Gruissem, B. and Freeling, M. 1995. Transposon-induced promoter scrambling: a mechanism for the evolution of new alleles. *Proc Natl Acad Sci U S A* **92:** 1836-1840.

Kuang, H., Padmanabhan, C., Li, F., Kamei, A., Bhaskar, P.B., Ouyang, S., Jiang, J., Buell, C.R., and Baker, B. 2009. Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: new functional implications for MITEs. *Genome Res* **19:** 42-56.

Lander, E.S. Linton, L.M. Birren, B. Nusbaum, C. Zody, M.C. Baldwin, J. Devon, K. Dewar, K. Doyle, M. FitzHugh, W. et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lieb, J.D. and Clarke, N.D. 2005. Control of transcription through intragenic patterns of nucleosome composition. *Cell* **123:** 1187-1190.

McClintock, B. 1947. Cytogenetic studies of maize and *Neurospora*. *Carnegie Inst Washington Year Book* **46:** 146-152.

McClintock, B. 1948. Mutable loci in maize. *Carnegie Inst Washington Year Book* **47:** 155–169.

Moreno-Vazquez, S., Ning, J., and Meyers, B.C. 2005. hATpin, a family of MITE-like hAT mobile elements conserved in diverse plant species that forms highly stable secondary structures. *Plant Mol Biol* **58:** 869-886.

Naito, K., Cho, E., Yang, G., Campbell, M.A., Yano, K., Okumoto, Y., Tanisaka, T., and Wessler, S.R. 2006. Dramatic amplification of a rice transposable element during recent domestication. *Proc Natl Acad Sci U S A* **103:** 17620-17625.

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C.N., Richardson, A.O., Okumoto, Y., Tanisaka, T., and Wessler, S.R. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* **461:** 1130-1134.

Oki, N., Yano, K., Okumoto, Y., Tsukiyama, T., Teraishi, M., and Tanisaka, T. 2008. A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, Oryza sativa ssp. japonica. *Genes Genet Syst* **83:** 321-329.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A. et al. 2009. The Sorghum bicolor genome and the diversification of grasses. *Nature* **457:** 551-556.

Rizzon, C., Martin, E., Marais, G., Duret, L., Segalat, L., and Biemont, C. 2003. Patterns of selection against transposons inferred from the distribution of Tc1, Tc3 and Tc5 insertions in the mut-7 line of the nematode Caenorhabditis elegans. *Genetics* **165:** 1127-1135.

Schnable, P.S. Ware, D. Fulton, R.S. Stein, J.C. Wei, F. Pasternak, S. Liang, C. Zhang, J. Fulton, L. Graves, T.A. et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326:** 1112-1115.

Spradling, A.C., Stern, D.M., Kiss, I., Roote, J., Laverty, T., and Rubin, G.M. 1995. Gene disruptions using P transposable elements: an integral component of the Drosophila genome project. *Proc Natl Acad Sci U S A* **92:** 10824-10830.

Tanaka, T., Antonio, B.A., Kikuchi, S., Matsumoto, T., Nagamura, Y., Numa, H., Sakai, H., Wu, J., Itoh, T., Sasaki, T. et al. 2008. The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res* **36:** D1028-1033.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O. et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8:** 973-982.

Yang, G., Nagel, D.H., Feschotte, C., Hancock, C.N., and Wessler, S.R. 2009. Tuned for transposition: molecular determinants underlying the hyperactivity of a Stowaway MITE. *Science* **325:** 1391-1394.

Zhang, Q., Arbuckle, J., and Wessler, S.R. 2000. Recent, extensive, and preferential insertion of members of the miniature inverted-repeat transposable element family Heartbreaker into genic regions of maize. *Proc Natl Acad Sci U S A* **97:** 1160-1165.

Zhang, X., Jiang, N., Feschotte, C., and Wessler, S.R. 2004. PIF- and Pong-like transposable elements: distribution, evolution and relationship with Tourist-like miniature inverted-repeat transposable elements. *Genetics* **166:** 971-986.

Chapter 5

Conclusions

Transposable elements (TEs) make up the majority of most higher eukaryotic genomes and have been believed serving as one of the strongest evolution forces (Dooner and Weil 2007; Feschotte *et al.* 2002; Feschotte and Pritham 2007). TEs can be divided into two main classes, class 1 (RNA) TEs and class 2 (DNA) TEs, based on the intermediate during the transposition process. Class 2 nonautonomous TEs, especially miniature inverted-repeat transposable elements (MITEs), are present at very high copy numbers in most sequenced genomes. Among cereal grasses, they are the predominant TEs that are in or near host genes (Naito *et al.* 2006; Naito *et al.* 2009; Spradling *et al.* 1995). The availability of four sequenced grass genomes, *Brachypodium distachyon*, *Oryza sativa* (rice), *Sorghum bicolor* (sorghum) and *Zea mays* ssp. *mays* L. (maize) provides a unique opportunity to characterize TEs and determine their genome wide association with grass genes (2005; 2010; Paterson *et al.* 2009; Schnable *et al.* 2009). However, currently available computational tools are limited at discovering and characterizing class 2 TEs. As such, I developed two programs, TARGeT and MITE-Hunter that can be used to discover both autonomous and nonautonomous class 2 TEs.

TARGeT (Tree Analysis of Related Genes and Transposons) is a homology based pipeline that can use either a DNA or a protein sequence as the query to identify, retrieve and characterize homologs from DNA sequence databases (see chapter 2) (Han *et al.* 2009). It is particularly useful if one wants to quickly retrieve and characterize homologs of a big gene or TE family from DNA databases. The core of TARGeT is a Perl program that automatically retrieves homolog sequences from the BLAST output. Once the homolog sequences are generated, TARGeT generates figures showing the gene structures and performs multiple alignment and phylogenetic analysis. TARGeT is both accurate and fast. TARGeT was evaluated using well annotated datasets, including the ascorbate peroxidase gene family of rice, maize and sorghum, and several TE families in rice. In all cases TARGeT accurately identified the known homologs and predicted new ones on the order of minutes. TARGeT is multifunctional. First, TARGeT has been used to discover autonomous DNA TEs. Second, the main component of TARGeT is also used in the MITE-Hunter pipeline (see chapter 3). Finally, TARGeT was used to discover domesticated transposase genes, which were not included in this dissertation.

MITE-Hunter is a program pipeline that can identify MITEs as well as other small class 2 nonautonomous TEs from genomic DNA datasets. MITE-Hunter was evaluated by applying it to the rice genome. The results were compared with known rice class 2 nonautonomous TEs in the Repbase (Jurka *et al.* 2005), whereas the TEs in Repbase were the products of many studies, MITE-Hunter was able to find virtually all the MITEs in a relatively short time frame and found eleven new families. MITE-Hunter was also

compared with two other MITE discovery programs, FINDMITE (Tu 2001) and MUST (Chen *et al.* 2009). Unlike MITE-Hunter, neither of these programs can search large genomic datasets including whole genome sequences. More importantly, MITE-Hunter is significantly more accurate than either FINDMITE or MUST as the vast majority of their outputs are false positives. The MITE-Hunter output is easy to curate as it contains highly condensed TE consensus sequences that are grouped into families. The validity of a TE discovered by MITE-Hunter can be quickly judged from the automatically generated MSA file, which is, to our knowledge, a unique feature of MITE-Hunter.

Using TARGeT and MITE-Hunter, I annotated and characterized class 2 nonautonomous TEs in *B. distachyon*, rice, sorghum and maize. Three features of class 2 nonautonomous TEs were revealed in this study. First, distinctive patterns of TE length and copy number were found for each superfamily. This result provides new insights into the origin and amplification of MITEs. Second, the association of class 2 nonautonomous TEs and MITEs with genes is described at high resolution in grass genomes. Of greatest interest is that MITEs cluster in the promoters of a large fraction of annotated genes, and are especially enriched within 1kb of the transcription start site. Putative mechanisms for the observed enrichment and the potential impact of MITEs on the evolution of gene expression are discussed. Finally, attempts were made to correlate the presence or absence of TEs in the 5' flanking regions of genes with different levels of gene expression. Interestingly, this analysis revealed that highly expressed genes were more likely to contain Tourist MITEs in their 5' flanking region than genes with lower levels of expression. Taken together, this study provides the most

comprehensive analysis to date of the association and potential impact of class 2 nonautonomous TEs and MITEs with grass genes.

References

2005. The map-based sequence of the rice genome. *Nature* **436:** 793-800.

2010. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* **463:** 763-768.

Chen, Y., Zhou, F., Li, G., and Xu, Y. 2009. MUST: a system for identification of miniature inverted-repeat transposable elements and applications to Anabaena variabilis and Haloquadratum walsbyi. *Gene* **436:** 1-7.

Dooner, H.K. and Weil, C.F. 2007. Give-and-take: interactions between DNA transposons and their host plant genomes. *Curr Opin Genet Dev* **17:** 486-492.

Feschotte, C., Jiang, N., and Wessler, S.R. 2002. Plant transposable elements: where genetics meets genomics. *Nat Rev Genet* **3:** 329-341.

Feschotte, C. and Pritham, E.J. 2007. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* **41:** 331-368.

Han, Y., Burnette, J.M., 3rd, and Wessler, S.R. 2009. TARGeT: a web-based pipeline for retrieving and characterizing gene and transposable element families from genomic sequences. *Nucleic Acids Res* **37:** e78.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110:** 462-467.

Naito, K., Cho, E., Yang, G., Campbell, M.A., Yano, K., Okumoto, Y., Tanisaka, T., and Wessler, S.R. 2006. Dramatic amplification of a rice transposable element during recent domestication. *Proc Natl Acad Sci U S A* **103:** 17620-17625.

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C.N., Richardson, A.O., Okumoto, Y., Tanisaka, T., and Wessler, S.R. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* **461:** 1130-1134.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A. et al. 2009. The Sorghum bicolor genome and the diversification of grasses. *Nature* **457:** 551-556.

Schnable, P.S. Ware, D. Fulton, R.S. Stein, J.C. Wei, F. Pasternak, S. Liang, C. Zhang, J. Fulton, L. Graves, T.A. et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326:** 1112-1115.

Spradling, A.C., Stern, D.M., Kiss, I., Roote, J., Laverty, T., and Rubin, G.M. 1995. Gene disruptions using P transposable elements: an integral component of the Drosophila genome project. *Proc Natl Acad Sci U S A* **92:** 10824-10830.

Tu, Z. 2001. Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, Anopheles gambiae. *Proc Natl Acad Sci U S A* **98:** 1699-1704.